Analysis and Modeling of U.S. Army Recruiting Markets

THESIS

MARCH 2016

Joshua L. McDonald, Major, USA

AFIT-ENC-MS-16-M-117

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENC-MS-16-M-117

ANALYSIS AND MODELING OF U.S. ARMY RECRUITING MARKETS

THESIS

Presented to the Faculty

Department of Mathematics and Statistics

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Joshua L. McDonald, B.S.

Major, USA

MARCH 2016

AFIT-ENC-MS-16-M-117

ANALYSIS AND MODELING OF U.S. ARMY RECRUITING MARKETS

THESIS

Joshua L. McDonald, B.S.
Major, USA

Committee Membership:

Dr. Edward D. White
Chair

Dr. Raymond R. Hill
Member

CPT Christian C. Pardo, USA
Member

AFIT-ENC-MS-16-M-117

Abstract

The United States Army Recruiting Command (USAREC) is charged with finding, engaging, and ultimately enlisting young Americans for service as Soldiers in the U.S. Army. USAREC must decide how to allocate monthly enlistment goals, by aptitude and education level, across its 38 subordinate recruiting battalions in order to maximize the number of enlistment contracts produced each year. In our research, we model the production of enlistment contracts as a function of recruiting supply and demand factors which vary over the recruiting battalion areas of responsibility. Using county-level data for the period of recruiting year (RY)2010 through RY2013 mapped to recruiting battalion areas, we find that a set of five variables along with categorical indicators for battalions and quarters of the fiscal year accounts for 70%, 74%, and 81% of the variation in contract production for high-aptitude high school seniors, high-aptitude high school graduates and all others, respectively. We find indications that high-aptitude seniors and graduates should be modeled as separate entities, contrary to current procedure. Finally, our models perform consistently well against a validation dataset from RY2014, and we ultimately achieve 530%, 119%, and 170% relative increases in respective correlation coefficients over previous comparable literature.

For Him Who is perfect in love and in Truth,

and without Whose redemptive power in my life all else is rubbish.

Acknowledgements

The completion of this project has been a joyfully arduous process. First and foremost, I would like to thank my family. By your steadfast love and support you gave me the freedom to enjoy doing this work.

I would next like to thank my advisor, Dr. Tony White. Sir, your professionalism and enthusiasm for the field are exceptional. Thank you for your time and attention to detail on the many drafts, spreadsheets, graphs, and other half-baked ideas that came your way. To my readers: Dr. Hill, thank you for taking on this paper in addition to the "few" other students you advised this year; to Christian, thank you for getting me started early-on with recruiting issues and for your tactical perspective. To Joe Baird and MAJ Mike Fleischmann at USAREC, thank you for your quick turnarounds on data requests and other support from the flagpole. I hope the utility of this project's results are worthy of the time you have invested in it, and in me.

To LTC Brian Lunday: Sir, thank you for your mentorship and guidance throughout my time at AFIT. Your dedication, professionalism, and thoughtfulness are unmatched. To Katie Timmerman of Wright State University, thank you for introducing me to the world of Java®, without which the data pull from the Census Bureau would have been impossible. Thank you to all those who assisted my editing; this is a much better effort because of your help. Finally, to JJ Dwyer and Rob Montgomery: you have been indispensable battle buddies in academics, life, and faith over the last eighteen months. Thank you for your friendship and for the privilege of knowing both you and your families.

Joshua L. McDonald

Table of Contents

Page

List of Figures

List of Tables

ANALYSIS AND MODELING OF U.S. ARMY RECRUITING MARKETS

I. Introduction

Since the formal elimination of the draft by Congress in 1973, the U.S. Army has maintained an All-Volunteer Force (AVF) [3]. Army Recruiters are tasked to help fill the ranks of the AVF by actively pursuing qualified future Soldiers with the ultimate goal of generating required enlistments. However, recent emerging trends present challenges to Army recruiters because the pool of potential Soldiers required to maintain the AVF appears to be decreasing. For example, only 3 in 10 American youth aged 17 to 24 years old are eligible for Army service, according to the U.S. Army Recruiting Command (USAREC) [4]. Increasing obesity, decreasing physical fitness, and decreasing reading ability are thought to be among prominent factors affecting decreasing service eligibility. Moreover, increasing attitudes of narcissism and decreasing propensity toward military service tend to further reduce the available pool of potential recruits [5]. And in 2015, the Army barely met its total recruiting goal only after sacrificing roughly two thousand of its Delayed Entry Pool (DEP) for the Reserve Component (RC) [6]. In light of this challenging environment, Army recruiting leadership requires increasingly accurate information regarding the market for its product: enlistment as an Army Soldier.

1.1 Research Purpose and Scope

The purpose of this research is to provide USAREC leadership with focused, relevant, and quantitative insight into its missioning process. The term *missioning* encompasses the process whereby Headquarters (HQ) USAREC decides how to dis-

tribute recruiting quotas to its subordinate units; a mission is the recruiting equivalent of a sales goal in the private sector. The missioning process results logically as recruiting leaders attempt to answer the question, "How does USAREC distribute its recruiting missions across the United States in a way that maximizes potential [enlistment] production [4]?" To effectively answer this question, USAREC must undertake at least two tasks in chronological order. First, USAREC must establish an accurate relationship between numerous recruiting factors—both within (i.e., demand) and outside of (i.e., supply) its control—and enlistment production in each geographical recruiting area. Assuming accurate relationships have been defined, USAREC must then set goals in a manner that takes advantage of these relationships to produce a maximum total number of projected enlistment contracts. We focus on the first of these tasks as it is fundamental to successful execution of the second. Also and for reasons which we detail in subsequent chapters, enlistment contract modeling efforts to-date leave considerable room for improvement.

Thus, we formulate our primary research question: *To what extent can we accurately express the relationship between enlistment supply and demand factors, and enlistment contract production*? To further focus our scope, we consider Regular Army (RA) enlistment contracts in the 50 States and the District of Columbia (D.C.). For added relevance, we ask the primary research question for each of USAREC's 38 recruiting battalions (i.e., recruiting markets) and three types of RA enlistment contracts.

By employing several mathematical methods, we ultimately ensure achievement of our research purpose by providing a quantitative answer to the primary research question. First, we take advantage of open source data at the county level; through a novel weighting technique, we use county-to-ZIP Code Tabulation Area (ZCTA) relationships to map county-level data to each battalion within USAREC. We initially

suggest a set of 26 candidate variables based on their ability to render complete situational understanding as defined by the Army's operational and mission variables [1]. We then apply ordinary least squares (OLS) mixed stepwise regression—aided by principal components analysis—to the candidate variables in order to estimate optimally-fitting, parsimonious models for each recruiting battalion and contract type. We use recent data from Recruiting Year (RY)2010 through RY 2013 to estimate these models. Finally, we validate our models by predicting contract production for a span of data not used in estimation, that of RY2014. In this last step we also create additional conditions of realism by using forecasts of the predictor variables. At the conclusion of this step, we achieve our penultimate objective by rendering quantitative battalion- and contract-specific comparisons of model performance within the context of an operationally relevant scenario. We now discuss the operationally pertinent aspects of USAREC's current missioning process.

1.2 Current USAREC Missioning Procedures

Overview and Terminology.

USAREC's analysis of recruiting markets and subsequent recruiting goal allocation decisions are collectively called Market and Mission Analysis (MMA). To conduct MMA, USAREC defines each recruiting market in terms of *elements* and *segments*. The four market elements are the [7]:

1. *Potential Market*: the proportion of the general population who show an interest in the Army, or would if they had better information;
2. *Qualified Military Available* (*QMA*) *Market*: the proportion of the potential market qualified for Army service and who are not currently serving in the military;
3. *Target Market*: males aged 17–24 who have a high school diploma and ≥ 50 (category I thru IIIA) Armed Services Vocational Aptitude Battery (ASVAB) test score;

4. *Penetrated Market*: the proportion of the potential market currently serving in the Army or who have enlisted and are awaiting basic combat training;

In addition to the market elements, USAREC uses market segments—individuals grouped by like characteristics—to further describe market conditions in order to more effectively apply specific recruiting strategies. For purposes of our research, the market consists of three mutually exclusive segments based on both education level and aptitude: Graduate Alphas (GA), Senior Alphas (SA), and Others (OTH). The definitions of these market segments are presented in Table 1 [4, 7].

Table 1. Recruiting Market Segmentation

Segment	Abbreviation	Education and Aptitude Criteria
Graduate Alpha	GA	A high school diploma graduate with $\geq$ 6 months since graduation, scoring in Test Score Category (TSC) I-IIIA (i.e, above the 50th percentile)
Senior Alpha	SA	A high school senior or diploma graduate within 6 months of graduation, TSC I-IIIA
Other	OTH	An individual not meeting the educational or aptitude criteria for GA and SA

As part of the federal budget process, HQ USAREC receives the *accession mission*: guidance from Department of the Army (DA) G-1 on exactly how many individuals must enter the Army during respective months of the fiscal year. The accession mission is specified by Army component and market segment. Following MMA, the five USAREC brigades are missioned; that is, they receive their respective portions of the *net contract mission*. The net contract mission consists of the accession mission corrected for anticipated breaches of enlistment contracts (known as "DEP-losses"), and including recommended contract missions for the next subordinate echelon (i.e., battalions). Factors considered in the assignment of the net contract mission include each battalion's past production, seasonal future losses, recruiter strengths, and geographical location with a goal of achieving equity in mission difficulty between battalions [7]. We now delve briefly into a key aspect of the missioning process which enables HQ USAREC's production of annual net contract missions.

The Recruiting Market Index (RMI).

The station is the lowest echelon to which a mission is formally assigned. We only model battalion-level missioning since the latter is the lowest echelon for which missions are generated by HQ USAREC.[1] Missioning for each year is currently a two-step process for HQ USAREC. In step one, each of the roughly 42,000 U.S. ZIP codes is assessed for potential production via a weighted combination of three factors over the previous four years: Production of all services, Army production, and the QMA population. At the conclusion of step one, ZIP code estimates are then weighted by another factor known as the Recruiting Market Index (RMI). The RMI is a linear regression conducted for each of USAREC's 38 battalions. The RMI response is a ratio of potential production to recruiter strength; a total of six predictor variables includes unemployment and historical productivity rates, among others [8].

Step two of missioning takes the RMI-weighted missions—which are still at the ZIP code level—and simply aggregates them to battalion and brigade echelons. In reality, the second step is slightly more complicated, since the RC mission is calculated and distributed at the ZIP code level prior to the Active Component (AC) mission. This intermediate step involves yet another set of weights on factors unique to the RC, but which is not relevant to the current research goal [8]. At this point we conclude our review of current missioning procedures.

1.3 Research Organization

We conclude the whole of our introductory material with an overview of the thesis layout. In Chapter II, we present a review of previous literature pertaining to recruiting market analysis. The review is by no means exhaustive, but does make a

[1]HQ USAREC only mandates brigade (BDE)-level missions. Battalion (BN)-level missions are formally recommended to each BDE HQ by USAREC, although ultimate authority for setting BN missions is delegated to each respective BDE HQ as the immediate commanding unit.

concerted attempt to introduce findings and methodologies representative of research in this area over the last nearly 30 years; supplementary material is located in Appendix B. In Chapter III, we introduce and develop the quantitative methodologies brought to bear on our research question; supplementary material for this chapter is provided in Appendices C through E. In Chapter IV we formally present our results and analysis, with supporting material in Appendix F. Finally, in Chapter V we revisit the original purpose and scope of the research, providing a concise comparison of our results with those of previous studies and current practices. We close our thesis with several recommendations for further study.

II. Literature Review

2.1 Introduction

The literature regarding enlistment behavior is abundant. We focus our discussion on studies in three general areas: macroeconomic enlistment models, microeconomic enlistment models, and choice analysis. In addition, we include a fourth category for previous research of potential value; we classify this category as simply "other research." We note that these broad categorical definitions are not mutually exclusive. However, some observed distinctions are helpful to negotiate the breadth of material available. For example, studies of enlistment supply at the macroeconomic level make use primarily of econometric regression models to estimate the effects of various supply and demand factors on the quality and quantity of enlistments. In general, these studies do not provide geographically specific observations or recommendations for recruiting resource allocation. Studies of enlistment supply at the microeconomic level extend the methodology of macroeconomic techniques to specific geographic locations and infer recruit production for areas as small as ZIP codes. The third broad category builds upon or otherwise employs concepts of discrete choice analysis as discussed by Ben-Akiva and Lerman [9]. These studies appear to rely more heavily on survey data and attempt to model behavior of specific recruiting market segments using multinomial probability models–either in lieu of, or in addition to an econometric specification. The miscellaneous category includes some qualitative studies and a goal program to determine optimal enlistment incentive allocation decisions. Table 2 provides a brief overview of the relevant research to be discussed in this chapter.

Table 2. Summary of Previous Research

Empirical Study	Broad Category	Methodology(ies)	Unit of Observation		
			Period Covered	Interval	Region
Dertouzos (1985)	Macroeconomic	Econometric (log-linear)	1980–1981	Month	Military Entrance Processing Station (MEPS)
Kilburn & Klerman (1999)	Discrete Choice	Multinomial logit	1994	–	National Educational Longitudinal Survey (NELS)
Murray & McDonald (1999)	Macroeconomic	Econometric (linear)	FY1983–FY1993	Month	Public Use Microdata Area (PUMA)**
Warner, Simon & Payne (2001)	Macroeconomic	Econometric	FY1988–FY1997	Month	County
Dertouzos & Garber (2006)	Macroeconomic	Linear regression; Logistic regression	Jan 2001-Jun 2003	Month	Recruiting station
Kleykamp (2006)	Macroeconomic	Logistic regression	FY2002	–	County
Dertouzos & Garber (2008)	Discrete Choice	Logistic regression	FY2001–FY2004	Month	Recruiting station
Asch, Heaton & Savych (2009)	Discrete Choice	Econometric	FY1998–FY2007	Quarter	State
Gibson, Luchman, Griepentrog & Marsh (2009)	Microeconomic	Zero-inflated Poisson regression; neural network; principal components analysis	FY2006–FY2008	Year	ZIP Code Tabulation Area (ZCTA)*
Gibson, Hermida, Luchman, Griepentrog, & Marsh (2011)	Microeconomic	Zero-inflated Poisson regression	FY2008–FY2009	Year	ZIP Code Tabulation Area (ZCTA)*

*Much of the market data come from state or county-level statistics which are subsequently appended to ZIP codes
**A PUMA is defined by the Census Bureau as a multi-county area

2.2 Macroeconomic Enlistment Supply

In 1985, Dertouzos conducted one of the first formal studies of enlistment supply to include demand factors (e.g., quotas, recruiter incentives, etc.) [10]. At the time, there existed little consensus over which factors actually impacted the supply of enlistments and to what degree; Dertouzos hypothesized that this may have been due to a previously incomplete formulation of the supply-demand relationship. He noted that "recruiters do not passively process enlistments; rather, by allocating their time differently in response to [missions] and to rewards for achieving and exceeding them, they alter both the quantity and quality of enlistments [10]."

Using data from 33 of 67 Military Entrance Processing Stations (MEPS) and years 1980 and 1981, Dertouzos proposed a regression model having a generally log-linear form [10]. The dependent variable was the number of high quality contracts (i.e., scoring in the top half of aptitude and having a high school diploma). Independent supply variables were the number of non high-quality (i.e., "other") contracts, unemployment rate, manufacturing wage, and the population of 15–19 year-olds. Independent demand variables included the number of recruiters and the contract missions for the

dichotomous enlistment quality stratification. Dertouzos accounted for both supply and demand factors simultaneously via a two-stage least squares (2SLS) estimation; he also estimated the model using OLS and maximum likelihood estimation (MLE), respectively, in order to compare results [10].

His results appear to confirm the hypothesis that demand factors have a significant impact on supply coefficient estimates at the 0.05 significance level. In particular, Dertouzos found that high-quality recruits would increase by 8.42% if the number of recruiters increased by 10% (all other factors held constant). This estimate was obtained from the 2SLS model, and was notably more conservative than either the OLS or MLE estimates at 9.61% and 11.9% percent, respectively. However, the 2SLS model was the only one to incorporate the full set of demand factors. Dertouzos also found high quality contracts to be four times more difficult to obtain than others, suggesting an explicit trade-off between efforts allocated to different quality categories [10].

In the early and mid-1990s, following the first Gulf War and dissolution of the Soviet Union, the U.S. Army experienced a significant reduction in fiscal and manpower recruiting resources. Failures of some recruiting stations to meet their missions and widespread reports of lowering propensity further fueled concerns about the future ability of the Army to meet its recruiting mission. In 1999, Murray and McDonald studied Fiscal Years (FY)1983–87 and FY1990–93 data to determine whether or not earlier models should have predicted such trends [11]. They used a linear specification of an econometric regression model to relate the number of high-quality, non-prior service contracts to a set of variables representing "youths' opportunities and the military's recruiting efforts." Their approach was derived loosely from earlier work by Dertouzos and by Dertouzos, Polich, and Press. Differences from the works of Dertouzos et al. were [11]:

- a linear model specification (since the logarithm of a zero-contract region is undefined), using feasible-generalized least square (FGLS) coefficient estimates; a logarithmic analysis is also reported, with the coefficients having a high degree of similarity to the linear specification.

- a differing assumption that the effect of goals on enlistments cannot be fully captured by the ratio of the latter to the former; therefore, the two variables are separated.

- the use of Public Use Microdata Areas (PUMAs) as the geographic data component; recruiting contracts were not reported by PUMA so they were obtained by a battalion-to-PUMA crosswalk.

Results of the study were mixed. In general, coefficients were lower in terms of effect than those reported in earlier studies, although significance was similar. The authors attribute much of the differences to the use of PUMAs: "the benefits of more appropriate geographically based measures may have been outweighed by the costs of greater measurement error [11]."

In 2001 Warner, Simon and Payne evaluated the effect of an expansion in the Navy College Fund (NCF) on Navy enlistment supply [12]. The Clemson University-based research team ultimately expanded their study of high-quality enlistment supply to four Services including the Army, Air Force, Navy, and Marine Corps. They postulated that "understanding the impact of changes in the economy, population, and recruiting programs on the supply of high-quality enlistments is needed to answer policy questions concerning the expansion of enlistment incentive programs since FY[19]94, including the NCF program [12]."

Warner et al. used monthly recruiting district (i.e., battalion)-level data, mapped by to counties in the 48 contiguous states, for each of the services spanning from

FY1988 to FY1997. They used a two-way fixed effects model (effects across states and time were assumed to be fixed), applied to panel data. They measured total contracts and total high-quality contracts as dependent variables; independent variables included socio-economic and demographic factors, incentives, and advertising levels, each scaled by the total respective population. Data for the independent variables were compiled from a variety of sources including Military Enlisted Processing Command (MEPCOM), respective Service databases, the U.S. Census Bureau's Current Population Survey (CPS), Department of Defense (DoD) military pay tables, and the Bureau of Labor Statistics (BLS). The authors calculated unemployment directly from raw estimates of employed and unemployed per state and month. Advertising data were obtained from P.E.P. Research, Inc. for expenditures, impressions, medium, month, county, and Service. Key results of this study are reported in Table 3 [12]. In addition to studying enlistment supply, Warner et al. also addressed propensity

Table 3. Impact of Various Factors on Army Enlistments in Warner et al. (2001)

(Source) Variable	Impact on High-Quality Enlistments (percent change)
(Demographic) 4 percent decrease in 35+ aged veteran population	−15
(Demographic) 11 percent increase in age group 17-21 college attendance	−11
(Socio-economic) 10 percent decrease in unemployment	−2 to −3.5
(Resource) 10 percent increase in Army recruiter strength	4 to 6
(Resource) 100 percent increase in (i.e., doubling) enlistment bonuses	12
(Resource) 10 percent increase in military pay	4 to 12
(Resource) 10 percent increase in Army advertising impressions	14

trends with data from the DoD's Youth Attitude Tracking Survey (YATS) over the years 1985–1998. Their research was consistent with earlier efforts, finding propensity to be "positively and significantly related to parents' past military service" for all major demographics at the 0.05 level of significance. However, there remained large unexplained variations in propensity over time [12].

In 2006, Dertouzos and Garber evaluated numbers and quality of Army enlistments for all months and recruiting stations between FY1998 and June 2003. Data

from FY1998 through FY2000 were incorporated into a multinomial logistic as well as a linear model, each with approximately 55 dependent variables reflecting various market qualities and recruiter characteristics. For data occurring between January 2001 and June 2003, they used a binary logistic regression model to predict the probability that a station achieved its high-quality mission. In this model, variables representing specific recruiter characteristics were replaced by reserve component attributes in order to investigate inter-component competition at the station level. Their results indicated the performance of both models to be nearly identical in terms of predictor significance: recruiter attributes were found to be less significant, in general, than market and mission factors. Specifically, the authors found that lower levels of recruiter effort were required to enlist quality youth in (listed in decreasing order of importance) [13]:

- low civilian-wage areas;
- areas where the QMA population is high relative to recruiter personnel strength;
- markets that are largely urban, have high non-Catholic Christian populations, and relatively low proportions of African-Americans and children living in poverty;
- the months of June, July, September and October (May is the worst);
- areas with high proportions of veterans less than 43 years of age and low proportions of veterans between ages 56 and 65; and
- any region that is not the Mountain region.

Overall, the study found that significantly less effort was required in recruiting markets where the preceding factors figured prominently. Hence, not all markets required the same effort levels [13]. Furthermore, assigned recruiting missions should have, but did not adequately account for such variations in effort [14]. The failure of

contemporary mission models to incorporate effort levels was a primary motivator of a follow-up study conducted by the same authors in 2008. In the 2008 study, Detouzos and Garber expanded their previous model to estimate enlistments as a function of recruiter effort, while adequately accounting for differences in market conditions [14]. Their primary research goal was to evaluate the utility of recruiting performance metrics in use by the Army at the time. Using monthly station-level data covering the period of FY2001–2004, the authors' investigation found generally that at intermediate levels of mission difficulty, recruiter effort increased as goal difficulty increased. However, this increase in effort also appeared to have a diminishing marginal return and may have even decreased in response to missions of extreme difficulty [14].

Dertouzos and Garber also made findings of relevance to individual market quality. For example, they noted that "market quality [was] an important determinant of recruiter effort levels [14]." Thus, market quality affects goal difficulty, which in turn affects effort. More specifically, they found market quality in a station's area of responsibility to be dependent on a myriad of factors; notable examples were QMA, ratios of youth to On Production Regular Army (OPRA) recruiters, demographic factors, and competition from other Armed Services. Several of these market factors represented an expansion of the original 55-variable model used in 2006 to a 68-variable model [14]. Perhaps the most salient outcome of the authors' 2008 research was their finding that the three separately missioned categories—GA, SA, and OTH—responded quite differently between markets [14]. Essentially, they concluded that any [reliable] metric used to evaluate recruiting performance "require[d] econometric analysis to estimate the difficulty of enlisting youth of different types in different local recruiting areas [14]."

Dertouzos and Garber also experimented with different time units of observation. They found that when the month was used as the time interval, proportions

of variation in production levels explained (i.e., R^2) were only 0.32, 0.10, and 0.27 for GA, SA, and OTH contract types, respectively. This relatively poor performance was improved by the greatest margin by aggregating the data over a period of six months; in this case, the proportions of explained variance improved to 0.65 for GA and OTH, and 0.31 for SA [14]. Unfortunately, the rather important and specific findings from both the 2006 and 2008 studies are tempered by the authors' omission of key statistical significance indicators (e.g., P-values) on the independent variables. And while we may have some idea of predictive capability based on the given R^2 values, Dertouzos and Garber did not use a validation dataset to evaluate this aspect directly.

In total, the studies of macroeconomic enlistment supply are helpful in describing the "big picture" of recruiting models. There seems to be some general agreement between these studies in the significance of a few select factors; unemployment, QMA population, and veteran population are three that come to the fore. However, the limits of such studies can also be seen in their limited capacity to predict expected recruit production for a specific geographical area. Microeconomic enlistment models attempt to do just that, as we review in the next section.

2.3 Microeconomic Enlistment Supply

In 2011 Gibson et al. predicted accessions for individual ZIP codes for each of the Armed Services with a ZIP Code Valuation Study (ZCVS) [15]. Gibson et al. utilized a zero-inflated Poisson regression model developed during a 2009 study effort by DoD Joint Advertising, Market Research & Studies (JAMRS) [16]. We use the authors' own language for a brief description of their methodology:

> "A zero-inflated Poisson model, unlike a standard Poisson (count) model, distinguishes between two processes causing an excessive number of zeros... [it] estimates two models, a count model and a logistic model, and

> combines them in the prediction of the outcome variable. Predictors are associated with either model, with count model variables predicting the number of accessions per ZIP code and logistic model variables predicting the number of ZIP codes with a count of zero accessions [15]."

A zero-inflated model appears to have some applicability due to the finding of Gibson et al. that nearly half of all [sampled] ZIP codes yielded zero recruits for a given year [15]. Additionally, the authors conducted a Vuong test to determine the superiority of the zero-inflated model over the standard Poisson model; a Vuong test evaluates the null hypothesis that competing models are equally close to the "true data generating process" against the alternate hypothesis that at least one model is closer [15, 17].

The authors gathered data from a wide array of government, publicly available, and proprietary sources. They used data from FY2008 and FY2009 to estimate accessions for each Service in each ZIP code in addition to evaluating significance of the independent predictors. Data not already available at the ZIP code (or ZCTA) level was calculated mostly at the state level and then appended to all ZIP codes within each respective state. A set of 55 independent variables spanned, generally, the variable categories defined by [18] but in more detail. However, Gibson et al. added several distance metrics, such as distances of each ZIP code to the nearest military installation, recruiting center, and university. Also, it is worth noting that the response variable in Gibson et al. (2011) was more limited in that it did not stratify contract qualities [15].

Table 4. Impact of Various Factors on Army Enlistments in Gibson et al. (2011)

(Source) Variable	Impact on Enlistments (percent change)
(Demographic) 10 percent increase in average age	−8.3
(Demographic) 10 percent increase in veteran population proportion	7.1
(Demographic) 10 percent increase in married population	5.5
(Socio-economic) 10 percent increase in property crimes	1.0
(Socio-economic) 10 percent increase in unemployment	1.3
(Qualification) 10 percent increase in English proficiency of multilingual students	0.9
(Resource) 10 percent increase in Army recruiter strength	1.1

Several Army-specific findings of Gibson et al. (2011) are captured in Table 4. In addition to Table 4, the West region yielded 37.6% more active duty Army recruits than did the Midwest and South; the Northeast lagged behind the Midwest and South by a further 18.5 percent. Every additional employee per business and every additional American College Test (ACT) score point were associated with 24% and 29.4% increases, respectively, in the odds of a ZIP code yielding zero recruits. Across the services, the number of recruiters appeared to have the greatest positive effect on accessions (interestingly, an increase in recruiter strength of any Service except the Marine Corps appeared to boost Army accessions). A second major factor was aggregate household income. Finally, a larger high school population—independent of population density—proved to be a significant predictor [15].

As part of an exploratory analysis, Gibson et al. also identify ZIP codes which differ significantly between actual and predicted accessions in 2010. Top under-performing ZIP codes (actual $<$ predicted) are those in El Paso, San Diego, and Los Angeles. Top over-performers (actual $>$ predicted) are Cumberland County (NC), Comanche County (OK), and Bell County (TX) [15]. Gibson et al. stopped short of highlighting, as we now note, that the three latter locations coincide directly with very large Army installations at Fort (Ft.) Bragg, Ft. Sill, and Ft. Hood, respectively.

Having now examined both macro- and micro-economic approaches, we acknowledge added value of the latter in its geographic specificity. Unfortunately, this enhancement appeared to come with a cost of reduced resolution in the response. Moreover, the work of Gibson et al. also lacks the use of a validation dataset. We explore a third modeling approach relating to individual enlistment decisions. Where macro- and microeconomic studies have thus far involved the collection of historical data at various levels of geographical aggregation, the next group of studies takes advantage of *survey data*, designed for and gathered from individual respondents.

2.4 Choice Theory

In 1999, Kilburn and Klerman studied the post-high school decisions of youth, as indicated by the 1992 and 1994 National Educational Longitudinal Survey (NELS). Kilburn and Klerman build on earlier decision-oriented based models of Hosek and Peterson [19, 20]. Hosek and Peterson modeled a dichotomous choice between enlisting or not enlisting; Kilburn and Klerman expanded this to three choices: enlist, attend college or work/other. Using 49 variables and a multinomial logit model, Kilburn and Klerman confirmed an earlier finding that graduates and seniors responded to different sets of factors. New findings concluded that a graduate with a parent [currently in] in the military significantly increased enlistment probability. For seniors, English as a second language significantly decreased enlistment probability [19].

Kleykamp used a multinomial choice model in 2006 to explore the post-high school decisions of a 2002 Texas high school graduating cohort [21]. A noted attribute of this study is its use of a survey sample following September 11, 2001 and the initiation of military action in Afghanistan. With a final sample size of 2,074 males and 15 independent variables, Kleykamp concluded the following:

- college aspirations increase the odds of choosing the military over work;
- military presence is significantly associated with enlistment among youth; the interaction of ethnicity and military presence is significant for Hispanic and other groups, but not African-Americans;
- for a 1 percent increase in the local military employment share, the odds of civilian employment or going to college are each reduced by 25 percent, relative to joining the military [21].

In 2008, Rostker et al. surveyed 5,373 new Army recruits at Basic Combat Training (BCT) and One-Station Unit Training (OSUT) locations [22]. Rostker et al.

focused on recruits aged 20 and older in order to determine what factors, if any, had influenced "later" enlistment beyond the period immediately coinciding with high school graduation. The impetus for the study was given by MEPCOM data which had indicated the fraction of recruits aged 20 and older increased from 35% of total recruits in FY1992 to about 56% in FY2008 [22]. Results indicated family influence to have a strong effect on the decision to join the military, regardless of the recruit's age; 83% of those surveyed had a close family Service-member. Furthermore, the proportion of new recruits with a father or mother in the military was over four times that of the general U.S. youth population. Also, 36% of older recruits reported there had been "no jobs at home" and 49% described any available jobs as having been "dead-end [22]."

Rostker et al. also found older recruits were about twice as likely as younger recruits to initiate contact with a recruiter, whether by phone or by mail. Older recruits were about 31% less likely to learn about recruiters from school, even though 55% choose some form of post-secondary education after high school. Thirty-eight percent indicated they simply "took time off" after high school. Rostker et al. used linear regression of dichotomous age categorical variables to analyze respective effects on promotion on retention. Their findings indicated both responses to be higher, in general, for older recruits than for younger recruits. However, this finding was sensitive to a specific combination of age and either promotion or retention, although the middle-range age groups of 22–24 and 25–27 showed the greatest overall increase in promotion and retention rates) [22].

Asch et al. conducted a 2009 study on enlistment choices of minorities in the Army and Navy [18]. A primary research question of their study was, "what factors affect[ed] the enlistment supply of different market segments to the Army... and how [did] these effects differ by market segment [18]." The authors utilized Army

enlistment data from FY1998 through FY2007 as well as demographic data from the CPS, in corresponding years. Dependent variables were high-quality contracts for White, African-American, and Hispanic enlistees, respectively. Asch et al. defined high-quality as a high-school diploma and an above-average Armed Forces Qualification Test (AFQT) score; it is not clear from their research if this definition included SA contracts, which may or may not possess a high-school diploma according to current USAREC definitions [7]. Thirteen independent variables captured market, mission, and demographic factors which parallel those chosen by previous research [18, 14]. Notable additions included obesity and crime rate, as well as the aggregation of age-specific veteran populations into a single demographic proportion [18]. Recruiting goals were included as a quadratic polynomial term in order to account for [unobservable] effort as a concave function of difficulty [14].

Table 5. Statistically Significant Results for the Army Enlistment Model of Asch et al. (2009)

Dependent Variable	Black	White	Hispanic
Log(bonus amount)	$p < 0.01$	$p < 0.01$	–
Log(recruiters/population)	$p < 0.01$	$p < 0.01$	$p < 0.01$
Log(military/civilian pay)	–	$p < 0.01$	$p < 0.05$
% receiving Army College Fund	–	–	$p < 0.05$
Iraq War Effect	$p < 0.01$	$p < 0.01$	$p < 0.10$
Presidential approval rating	–	–	$p < 0.01$
Log(unemployment rate)	–	$p < 0.01$	–
Log(% veteran)	–	–	$p < 0.01$
Log(% non-citizen)	$p < 0.10$	$p < 0.05$	–
Log(% obese)	–	$p < 0.10$	–
% enrolled in college	–	$p < 0.10$	–
Log(crime rate)	$p < 0.01$	$p < 0.05$	–

The data was further organized by quarter and by state, and modeled using econometric panel data regression. The results are indicated in Table 5. The only independent variable not shown in Table 5 is the Montgomery GI Bill (MGIB) benefit; it was found to be insignificant for all groups. In the table, a "–" indicates insignificance (i.e., a P-value greater than 0.10). Table 5 summarizes the conclusion that demographics

responded differently to market factors [18]. It appears that recruiter-to-population ratios and the effect of the Iraq war were significant to all three demographics.

At this juncture, we conclude our review of the three major types of studies. Clearly, there is some overlap between the approaches we refer to as "choice theory" and those of the macro- or microeconomic nature; methodologies of panel data regression and logistic regression are common to all three categories. We now turn quickly to a few more studies which are of additional use, but which are best collectively characterized by their differences from each other, as well as from the studies examined thus far.

2.5 Other Research

In 2001, Henry et al. formulated and implemented a binary integer goal program to meet USAREC Mission Occupational Specialty (MOS)-specific recruiting goals subject to budgetary constraints [23]. Their model consisted of approximately 64,000 decision variables indicating which types of enlistment incentives to offer each prospective MOS. While the study did not explore the impact of market demographics, it did provide an approach for optimizing recruiting resources, provided appropriate input probabilities could be established. Additional mention is given to the use of choice analysis similar to those studies discussed in the previous section [18, 23].

Bicksler and Nolan comprehensively reviewed Joint Service recruiting studies through 2009 [24]. Several of the studies already detailed in this document constitute significant portions of their source material. One of their unique observations comes from polling data. In 2009, 82% of the American public had high confidence in the military as an institution, but this did not necessarily translate into propensity. For the same poll period (2009), propensity was about 15%, down from 26% in 1989. The authors of [24] assert that "given the established link between propensity and

enlistment, this long-term decline in propensity is significant, and presents serious challenges to today's military recruiters [24]." They highlight several other studies which make respective notes of:

- a dramatic projected increase in the Hispanic population, from 20% of youth in 2010 to 38% by 2050;[1]
- the allocation-breakdown of fiscal recruiting resources for the DoD;[2]
- the presence of a lag between spending on advertising and incentives and cyclic fluctuations in enlistments;
- a summary of relative effects of changes in recruiting resources, as indicated by Table 6; the most effective resource for boosting high-quality recruits—military pay—is also the most expensive (with a marginal cost of $200,000 per recruit, based on a 4-year enlistment).

Table 6. Impact of Various Factors on Army Enlistments, Bicksler and Nolan (2009)

(Source) Variable	Impact on Enlistments (percent change)
(Resource) 10 percent increase in recruiters	4.1 to 4.7
(Resource) 10 percent decrease in recruiters	−5.6 to −6.2
(Resource) 10 percent increase in advertising budget	−1.0
(Resource) 10 percent increase in bonus amount	0.5 to 1.7
(Resource) 10 percent increase in military pay	7.0 to 11.3
(Market) 10 percent increase in unemployment	2.0 to 4.0
(Market) War in Iraq	−12 to −33

Lastly, it is useful to review key points from USAREC's own doctrine concerning the importance of recruiting market factors. USAREC Manual 3–03 is the primary

[1] As a group, Hispanics have been historically predisposed toward military service but are underrepresented by about 3% in the U.S. Military [18, 24].

[2] In FY2008, 30% of Joint recruiting dollars was allocated to "field recruiters and supporting manpower." A further 23% was allocated to functions in direct support of recruiters (automation, logistics, etc.) with the remaining 19% and 24% dedicated to advertising and incentives, respectively [24]. Whether or not this distribution is uniform across all services is unclear.

doctrinal publication which prescribes operating guidance for recruiting brigades and battalions [7]. The manual cites unnamed studies which have purportedly demonstrated several trends. First, political factors influence recruiting (e.g., upcoming national elections may cause youth to postpone an enlistment decision and enlistments are positively correlated with elected officials' attitudes toward military service). Next, the close proximity of active military installations tends to increase enlistments. Also, recruiting stations of other Joint Services tend to decrease Army enlistments. It is further stated that unemployment rates and enlistments are positively correlated. Finally, economically depressed areas have higher enlistment rates [7].

2.6 Conclusion

We have now reviewed available and pertinent literature on Armed Forces recruiting covering the period from 1985 through 2011. Over that 26-year span, we have seen commonalities and differences between literature objectives, methods, and results. Reporting each study's results in a single table, for the purpose of making broad comparisons, might seem exceedingly useful. However, we are cautious that differing conditions and objectives between studies—however subtle—may lead to erroneous interpretations. As a compromise, in Appendix B we offer a comprehensive list of independent and dependent variables used in each study we reviewed. Reading down the variable name column, it is easy to see how apparently identical variables contain important differences.

We can draw some additional conclusions from prior literature. First, the literature we reviewed was dominated by econometric methodology. The econometric studies captured here shared a common objective to describe socioeconomic effects on recruiting over time. In nearly every study, this was accomplished by some form of

regression. We acknowledge the utility of regression techniques and propose a similar methodology in Chapter III.

Second, there appears to be some broad agreement that several factors are correlated with recruit production. As examples, we note specifically unemployment, veteran population, age demographics, recruiter strength, and monetary incentives (of these, unemployment has generally been found to have less relative importance than its counterparts). Unfortunately, we cannot conclude exactly how the magnitude of these factors' effects changes with geographic location. The different geographies used in each study make this task difficult; we also recognize that statistics at smaller geographies may have greater measurement error [11]. On the other hand, we have reason to think that geography is important; this assertion is based on the total body of empirical results, as well as the fact that USAREC allocates its recruiting missions by geographic boundaries in the first place [13, 14]. Unfortunately, a gap has emerged in the fact that no study contained results which were aggregated or reported according to specific recruiting unit boundaries. Therefore, the literature gives us a starting point for variable selection (see Appendix B) while allowing us to fill a knowledge gap by characterizing the effects of these variables in regions of operational significance to USAREC.

Finally, we note that virtually no space in previous research is devoted to specifying predictive models. By predictive models, we mean a type of model that is designed to produce forecasts into future time periods. Recruiting—like any private-sector marketing effort—requires decision-making (i.e., an irrevocable allocation of resources) in the face of uncertainty [25]. While the studies we reviewed provided some indication of how variables respond to time, most did not explicitly describe the response of a variable in "future" time or provide any kind of probabilistic statement regarding such future behavior. We shall use this observation as a primary motivator

for our methodology, specifically with regard to validation efforts.

Overall, we have found the body of literature surrounding Armed Forces recruiting to be fairly substantial and we have chronicled a relatively small portion of that research here. Nevertheless, we have sought to provide a representative sample that will inform the subsequent methodology, results, and conclusions of our own study. It is the methodology of our original efforts to which we now turn.

III. Methodology

3.1 Introduction

Our analysis involves a wide array of techniques including data gathering, imputation, variance reduction, mixed stepwise regression, and multiple linear regression. These techniques act in support of one another to form a comprehensive analysis picture. In this chapter, we describe each technique in a logical order, but our discussions of analysis and results in Chapter IV will not necessarily conform to this order.

Thus, we begin in Section 3.2 by providing a doctrinal framework to assist with our initial data selection. In Section 3.3 we describe each data source in greater detail, as well as any required data cleaning (e.g., imputation). Here we organize our data descriptions with the aid of the doctrinal framework initially presented. We conclude Section 3.3 with a short discussion of the database structure, in preparation for the main body of our methodology presentation.

Sections 3.4–3.7 discuss in detail the mathematical underpinnings of our analysis. We begin by discussing a useful variance reduction method, principal components analysis (PCA), in Section 3.4. In Section 3.5 we introduce OLS regression, our chosen method for mathematically modeling recruiting contract production. In Section 3.6 we build on the OLS discussion and introduce mixed stepwise regression, which is useful in obtaining parsimonious models with superior explanatory capability. Finally, in Section 3.7 we describe a strategy for testing the obtained regression models against new data to assess their overall utility. We close this final section with a brief summary that helps guide our transition into the presentation of results, in the next chapter.

3.2 Data Gathering

Our over-arching goal is to develop an adequate mathematical model which predicts recruiting production for a unit by looking at observable factors within the unit's area of operation. Hence, we must gather two broad sets of variables. The first set describes what is to be predicted (i.e., dependent variables). We define the dependent variable initially as some number and type of recruiting contracts. The second set of variables describes those observable conditions of the recruiting market/mission which ostensibly affect the outcome of the dependent variable. This second variable set is independent; that is, we assume these variables to be unaffected by the dependent variable or by each other.

In gathering our variables, we were immediately confronted by a fundamental difficulty stemming from the cross-sectional nature of the data. Recruiting data provided by USAREC exists at the battalion level—an aggregation of ZIP codes—and sampled at monthly intervals. However, data describing market conditions within each battalion is reliably and consistently available only down to the county level, sampled at annual intervals. Therefore, some way of mapping one entity's observational units to the other is required. It appears previous literature has either approximated unit boundaries to conform to standard political borders, or not addressed the incongruence altogether. These approaches may have been appropriate within the context of their respective research goals, but are not sufficient to address USAREC's current need of market-specific predictive accuracy.

Therefore, we propose the following two-step process as a solution. In Step 1, we gather annual county-level data (where possible) and subsequently weight this data—through a series of crosswalks using proportions of the general population—to ZCTAs. At the conclusion of Step 1, we aggregate the weighted ZCTA data to the recruiting battalion level. In Step 2, we interpolate monthly values between each of the annual

battalion data points. Thus, the entire process brings annual county-level data into conformity with the monthly battalion-level data structure provided by USAREC. We address the entire procedure in more detail shortly, with supplementary material provided in Appendices C and E.

Variable	*Description*
Political	Describes the distribution of responsibility and power at all levels of governance—formally constituted authorities, as well as informal or covert political powers
Military	Explores the military and paramilitary capabilities of all relevant actors (enemy, friendly, and neutral) in a given operational environment
Economic	Encompasses individual and group behaviors related to producing, distributing, and consuming resources
Social	Describes the cultural, religious, and ethnic makeup within an operational environment and the beliefs, values, customs, and behaviors of society members
Information	Describes the nature, scope, characteristics, and effects of individuals, organizations, and systems that collect, process, disseminate, or act on information
Infrastructure	Is composed of the basic facilities, services, and installations needed for the functioning of a community or society
Physical environment	Includes the geography and manmade structures, as well as the climate and weather in the area of operations
Time	Describes the timing and duration of activities, events, or conditions within an operational environment, as well as how the timing and duration are perceived by various actors in the operational environment

Figure 1. The Operational Variables, *Army Doctrine Reference Publication* 5-0 [1]

As chronicled in Appendix B, our review of previous literature revealed roughly 200 variables thought to characterize recruiting markets. Amidst project time and resource constraints, amassing this many metrics is infeasible. On the other hand, we cannot arbitrarily choose variables since doing so may omit potentially important aspects of recruiting behavior. As a solution, we apply the eight operational variables, known commonly as "PMESII-PT," to help focus our data gathering efforts. We also utilize the five mission variables, known by the pneumonic device "METT-TC." Army leaders define the operational and mission variables to increase situational understanding in full spectrum operations [1]. Though traditionally applied within a strict military context, we find the operational and mission variables suitable for describing recruiting conditions within the U.S.; in fact, USAREC cites a form of the operational variables in its own doctrinal literature [7]. We provide a summary of each of the main operational variables in Figure 1. Each of the operational variables

in Figure 1 can be further divided into several sub-variables, for a total of 48 possible metrics. The operational variables and sub-variables describe salient aspects of market conditions, which can also be thought of as recruiting supply factors since they are mostly external to USAREC's control.

Likewise, the five doctrinal mission variables—*mission, enemy, troops, terrain, time, civilian considerations*—describe conditions of the recruiters themselves. These can be interpreted as recruiting demand factors, since the recruiters act like salesmen to generate demand for the Army profession according to specific marketing strategies employed by USAREC. A complete crosswalk of the operational sub-variables and mission variables with our selected metrics is given in Figure 2. Our strategy was to

Variable Type	Variable Name	Sub-variable Name	Metric		
			Name	z_j	Definition (Time Unit of Measure and Geography)
Operational (Recruiting Supply)	Political	Government effectiveness & legitimacy	Voter Participation Rate	1	votes cast for President / total adult population (2008 and 2012, County)
	Military	Military forces	Sponsor Share	2	number of Army active duty sponsors / total active duty military sponsors (2010—2013, Annual, ZIP code)
	Economic	Economic activity	Labor Participation Rate	3	persons in labor force / total working-age population (2010—2014, Annual, County)
		Employment status	Unemployment Rate	4	employed persons / persons in labor force (2010—2014, Monthly, County)
	Social	Education level	Cohort HS Graduation Rate	5	graduates from freshman high school class / size of freshman class (2010—2014, Annual, County)
		Criminal activity	Violent Crimes	6	number of violent crimes (2010—2014, Annual, County)
		Basic cultural norms and values	Obesity	7	number of obese persons / total population (2010—2014, Annual, County)
			Illicit Drug Use	8	number of persons using illicit drugs / total population (2010 and 2012, County)
	Infrastructure	Urban zones	Urban Population Rate	9	number of persons in urban zones / total population (2006 and 2013, County)
	Information	Intelligence	Propensity	10	number of youth inclined toward military service (2010—2014, Semiannual, Battalion)
			QMA Population	11	number of youth aged 17—24, qualified without a waiver (2010—2014, Annual, ZIP Code)
			17-24 Population	12	number of youth aged 17—24 (2010—2014, Annual, ZIP Code)
	Physical	Terrain	Battalion Recruiting Station Identifier (RSID)	13	recruiting battalion boundaries (2010—2014, Annual, ZIP Code)
	Time	Information offset	Lag-1	14	number of total contracts produced from previous month (2010—2014, Monthly, Battalion)
Mission (Recruiting Demand)	Mission		Reg. Army GA Mission	15	goal for number of GA contracts (2010—2014, Monthly, Battalion)
			Reg. Army SA Mission	16	goal for number of SA contracts (2010—2014, Monthly, Battalion)
			Reg. Army OTH Mission	17	goal for number of OTH contracts (2010—2014, Monthly, Battalion)
			Reg. Army GA Achieved	18	number of adjusted GA contracts produced (2010—2014, Monthly, Battalion)
			Reg. Army SA Achieved	19	number of adjusted SA contracts produced (2010—2014, Monthly, Battalion)
			Reg. Army OTH Achieved	20	number of adjusted OTH contracts produced (2010—2014, Monthly, Battalion)
	Enemy (i.e., Competitors)		Contract Share	21	number of Army contracts / all DoD contracts (2010—2014, Monthly, Battalion)
			Recruiter Share	22	number of Army recruiters / all DoD recruiters (2010—2014, Monthly, Battalion)
	Troops and support available		Army Recruiters	23	number of Army active and reserve recruiters based on PERSTAT (2010—2014, Monthly, Battalion)
			Appointments Made	24	number of appointments scheduled and reported to USAREC (2010—2014, Monthly, Battalion)
			Appointments Conducted	25	number of appointments conducted and reported to USAREC (2010—2014, Monthly, Battalion)
	Terrain and Weather		See *Physical* Operational Variable		
	Time		Processing Days	26	number of days to process recruits (2010—2014, Monthly, Battalion)
	Civil considerations		See *Information-Intelligence* Operational Sub-variable		

Figure 2. Variable-to-Metric Crosswalk

include at least one sub-variable and metric for every operational variable, and to this end we were successful. However, we were only able to feasibly obtain data metrics for 11 of the 48 total operational sub-variables. So while our selected data renders a general situational picture, considerable knowledge gaps remain. In the next section, we briefly describe each of the metrics in Figure 2. The operational variables required a majority of the data pre-processing, so we begin there following a short discussion of notation.

3.3 Data Description

Before discussing the data in detail, we pause to introduce some brief notation conventions. These will be useful in subsequent sections where we desire brevity in referencing individual variables. First, we use y to denote dependent variables. We denote independent variables with x. We use z generically where separate role distinctions are not necessary. We also use a convention of sub- and super-scripted indicies to denote additional distinctions of z as required. This convention follows the general form

$$z_{j,t}^{(k,i)} \tag{1}$$

where

$i \equiv$ the index of the battalion RSID[1]; $i =$1B,1D,...,6N

$j \equiv$ the index of the variable from Figure 2[2]; $j = 1, 2, \ldots, 33$

$k \equiv$ the index of the contract type; $k =$ GA, SA, OTH

$t \equiv$ the index of the the observational unit time; $t = 1, 2, \ldots, 60$

[1]A complete list of Recruiting Station Identifications (RSIDs) and corresponding geographic locations is located in Appendix A.

[2]For reasons that will become apparent in Chapter IV, we will add several variables to the current maximum index of $j = 26$ as indicated by Figure 2.

When a definition applies to all elements of an index, we omit that index for brevity. However, we do include all relevant indicies when a definition or operation as individual needs for specificity dictate. Having completed this clerical note, we now move into specific descriptions of the variable data, beginning with the operational variables.

Operational Variables.

We collected a majority of our data on operational variables from open sources. We define an open source as a source which is available to the general public or to properly credentialed DoD personnel through limited-access portals such as the Defense Manpower Data Center (DMDC). As previously mentioned, much of our open source data is at the county level. Because of this, we incur the need for alignment of county geography with the ZIP code geography used by recruiting battalions. Our proposed method for this is a population-based weighting of each county-level datapoint to ZCTAs, followed by aggregation of the resulting ZCTA datapoints to a battalion-specific value. We use ZCTAs because their boundaries are much more consistent over time than those of ZIP codes [26]. However, this adds additional complexity because ZIP codes must, in turn, be "crosswalked" to ZCTAs due to overlaps. We provide a detailed explanation of our procedure to align ZCTAs and ZIP codes in Appendix C. Presently, we give the basic mathematical formulation of our county-to-ZCTA weighting technique, given that we have already cross-walked ZIP codes to ZCTAs.

Let $Z_i \subseteq Z$ be the subset of $(m = 1, 2, \ldots, 32846)$ ZCTAs within each unit i boundary. Let C be the set of $(n = 1, 2, \ldots, 3141)$ counties in the United States. Let C_i' be the set of counties which intersect with a ZCTA in a unit's area of responsibility $(C_i' \subseteq \{C \cap Z_i\})$. We then define a weighted statistic z', for unit i (scripts j and t are

omitted since this definition applies to all variables and times) as

$$z_i' = \sum_{m \in Z_i} \sum_{n \in C_i'} \upsilon_{m(n)} z_n \tag{2}$$

where

$\upsilon_{m(n)} \equiv$ the proportion of county n population residing in ZCTA m, from the 2010 Census [27, 28]

$z_n \equiv$ the available statistic for county n, where $|z_n| \geq 1$

An example illustration of the complex overlapping nature of battalion and county boundaries is given in Figure 3. Often the desired value of z_i is a rate (e.g., unemployment), in which case we apply (2) separately to the numerator and the denominator prior to dividing. In fact, we used (2) only for fractional data. We explored weighting a raw value such as population, but found that aggregating to unit levels produced a total value greater than the original. This is likely due to some double-counting in our formulation of z_i'. However, similar over-estimation errors applied to the numerator and denominator of a single rate are likely to be negligible in the end. The overall

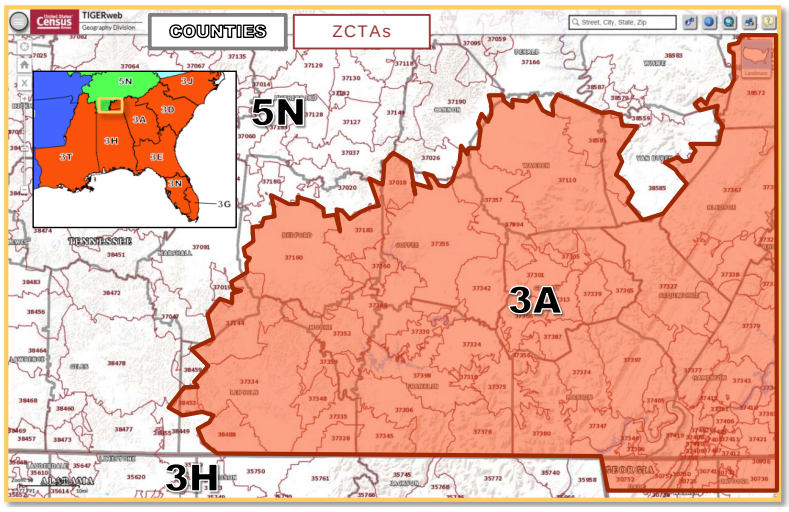


Figure 3. Portion of the Boundary for Battalion 3A (Atlanta) Showing ZCTAs and County Overlaps

reasonability of our resulting weighted values further increased our confidence in this method.

Much of our open source data also required imputation to replace missing values. Some data arrived incomplete at the county-level, but nearly all required imputation of monthly values from annual samples. In the latter and most frequent case, we used a technique known as stochastic mean value imputation. This is a variation of the mean-value method, but adds a random variable to the mean value to capture added variability [29]. To illustrate our implementation, let z_t and z_{t+12} be realizations of a battalion-level statistic at the same month in subsequent years, where the in-between monthly values of $z_{t+1}, z_{t+2}, \ldots, z_{t+11}$ must be imputed. We obtain the mean values for all imputed t by subtracting z_t from z_{t+12} and dividing by 12 to obtain the gradient, δ. Then we have the means $\hat{\mu}_t = z_t + \delta t$ for $t = 1, 2, \ldots, 12$. The standard deviation $\hat{\sigma}$ is then $(12\delta)/4 = 3\delta$ since by the empirical rule approximately 95% of the data lies within $\pm 2\sigma$ [30]. At this point we now have the two parameters, $\hat{\mu}_t$ and $\hat{\sigma}$, which characterize a normal distribution based on sample data. We then use the computationally straight-forward inverse transform technique to compute random realizations of this normal distribution for each time t, bearing in mind the lack of a closed-form inverse solution to the normal distribution necessitates a numerical computation [31]. We utilized the `norminv` function of Excel®2010 to perform the inverse transformations, supplemented with Visual Basic Application (VBA) code which is given in complete form in Appendix E. For brevity, the basic process is illustrated in Figure 4.

In cases where data was missing from the original county-level datasets, we frequently used "hot-deck" imputation. Hot deck imputation consists of essentially imputing a missing value in a given category from an existing observation in a similar category [29]. We took care to apply this method to areas of geographic similar-

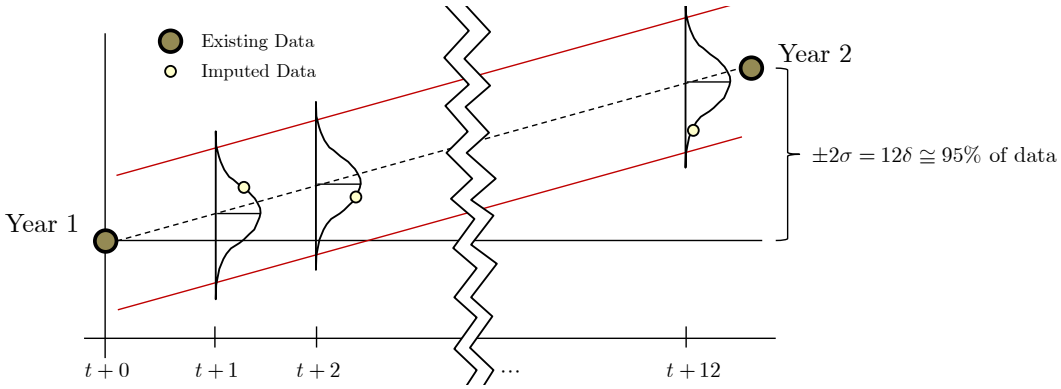


Figure 4. Illustration of Stochastic Mean Value Imputation

ity (i.e., using a value from Montana to replace missing values from Wyoming, not Florida). In several cases, the data required use of both the hot-deck method for imputation of annual data followed by the stochastic mean value method for imputation of monthly data. At this point we conclude our discussion of data cleaning techniques and move forward with descriptions of the individual datasets themselves.

Political Variable Metrics. Previous literature contained such political metrics as Presidential approval ratings or polls of public opinion on specific policies. We found these specific metrics to be difficult to locate over consistent time periods and geography. However, we did locate county-level voting statistics for the Presidential elections of 2008 and 2012. We refrained from a political party-oriented metric due to the potential for controversy. However, the voter participation rate in the general elections seems appropriately neutral to link with the operation sub-variable, "government effectiveness and legitimacy." We define the voter participation rate, z_1, as the total votes cast for President divided by the voting age population. The data sources are *The Guardian* and the U.S. Census Bureau American Community Survey (ACS), respectively [32, 33, 34]. This dataset required both hot deck and mean value imputation, as well as ZCTA-weighting.

Military Variable Metrics. Previous literature of the microeconomic type included several metrics to describe the geographic proximity of each market to military installations. We were not able to replicate this with a distance metric, but we suggest an alternative called "Sponsor share" (z_2): population of Army Service-members relative to total DoD Service-members. This is our attempt at expressing the public's exposure to military presence in their communities. We obtained the number of Active Duty Service-members for each major U.S. installation, from 2010 to 2013 [35].[3] ZIP codes included with each installation allowed us to forgo ZCTA-weighting. Hot deck and mean value imputation methods were applied.

Economic Variable Metrics. We include the labor participation rate (z_3) and unemployment rate (z_4) as metrics of the economic sub-variables employment status and economic activity, respectively. For reference we present our weighted BDE unemployment rates in Figure 5. The current USAREC missioning model incorpo-

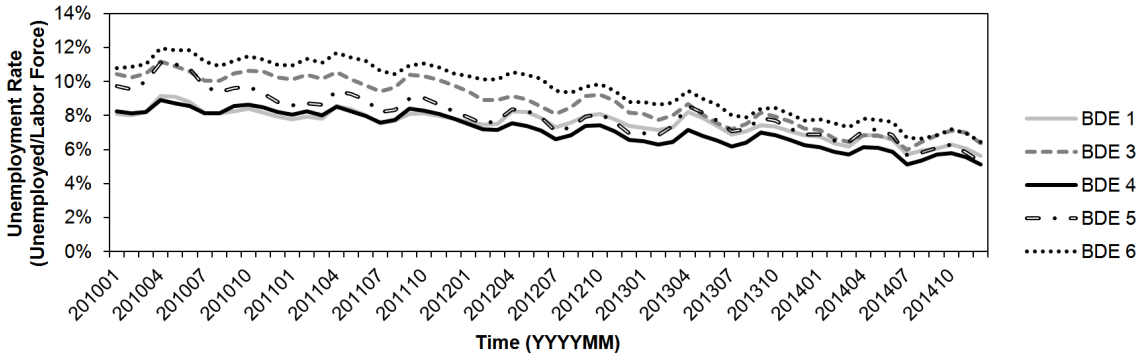


Figure 5. Unemployment Rate Using Weighted County Data (x_9) by Brigade RSID, FY2010–FY2014

rates unemployment as an independent variable. For the number of unemployed we use county-level, monthly data from the BLS [36]. The labor participation rate was not available from BLS at the county level, so we obtained this data from the 5-year

[3]The DMDC officially provides this data by installation on a monthly basis, but our time constraints did not allow for appropriate formatting of their online query-generated reports.

ACS [34]. Additionally, labor force size required stochastic mean value imputation since it was only available annually. The estimates in Figure 5 are broadly consistent recent levels and trends of decreasing unemployment which, as we mentioned previously, lends some credence to the geographic weighting technique we employed.

Social Variable Metrics. All four of our social variable datasets were obtained from The County Health Rankings and Roadmaps Program [37]. We selected the high school graduation rate (z_5), number of violent crimes (z_6), adult obesity rate (z_7), and the illicit drug use rate (z_8) to represent corresponding sub-variables. Previous literature included several or all of these variables in some form. Since all four datasets are similar in structure, we discuss them together. The original data was available at the county and annual intervals for 2010 to 2014. The variables z_5, z_7, and z_8 required conversion of original percents back to integral population numbers using an accompanying population. The weighting scheme was then applied to numerator and population separately, and after dividing we arrived back at the appropriate percentage expressions. Hot deck and stochastic mean value imputation methods were then applied.

Infrastructure Variable Metrics. We use a single metric for infrastructure, which we define as the "urban population rate" (z_9). We define the urban population rate as the percentage of persons living in large central metropolitan counties down to medium metro counties, relative to the total population. To obtain this metric, we use the population data from the 5-year ACS and a urban-rural county classification scheme provided by the Centers for Disease Control and Prevention (CDC) [38].[4] The urban and rural classifications were only provided for 2006 and

[4]The U.S. Census Bureau only provides an urban-rural classification for counties *within* Core-Based Statistical Areas (CBSAs). Since CBSAs exclude less populous areas by definition, the list of U.S. counties used in the Census Bureau's urban-rural classification is not collectively exhaustive.

2013. Thus, we applied the 2006 classification to the 2010, 2011, and 2012 population data from the ACS, with the 2013 classification being applied to the other two years. Finally, we applied ZCTA-weighting and stochastic mean value imputation.

Information Variable Metrics. We discuss three metrics that reflect the information operational variable. We define information within a more specific context of intelligence, meaning the specific knowledge recruiters have about their markets. We begin by defining a metric called "propensity" (z_{10}). Propensity is the fraction of "definitely" and "probably" responses of youth in a semi-annual DoD poll aged 17–24 to the question, "How likely [is it] that you will be serving in the Military in the next few years [39]?" USAREC provided this data to us at the battalion level, so only mean value imputation was required. In this case, we utilized a uniform distribution of the stated margin of error of $\pm 3\%$ to impute the random realizations, although this too is a greatly simplifying assumption.[5] Two other components of recruiter intelligence are the QMA population (x_{11}) and the 17–24 population (x_{12}). These data were provided by USAREC at annual intervals and the ZIP code level, so only mean value imputation was required. The data is calculated by a private firm, Woods & Poole Economics. The QMA population conforms to the definition provided in Chapter I, while the 17–24 population is self-evident. We include these metrics based on previous literature, as well as their prominence in USAREC's missioning decisions.

Physical and Time Variable Metrics. We conclude our description of operational variables with a brief mention of geographical and time-related aspects. Aside from the battalion as our geographical unit of measure, we do not include a

[5]The margin of error in the Youth Poll results from a sample in nine census-based regions, not the 38 battalion areas. In actuality, the battalion margin of error is likely higher than $\pm 3\%$ but we use the reported margin in the absence of better information.

separate metric for terrain. We assume that the battalion is an adequate level of geography about which inferences can be made regarding homogeneity of the market. In other words, effects of certain variables over time are likely to be relatively similar *within* each battalion, although they may be quite different *between* battalions. This is evident upon visual inspection of a few of the time series variables already discussed; see Appendix D. Regarding time, we do acknowledge the important role played by lagged responses in model formulation. Currently, USAREC incorporates lagged responses over previous years to aid its missioning process [8]. Exactly how we incorporate lagged response values is the subject of a later section in this chapter. We now leave the operational (supply-side) realm and turn to a description of our mission (demand-side) variables.

Mission Variables.

As we have previously mentioned, mission metrics are useful in characterizing the goals and performance of each recruiting battalion. The mission set of variables comes primarily from USAREC's own database. The availability of this data was a driving factor for how we gathered the operational variables, and we requested mission data for the period of FY2010 through FY2014 for two reasons. First, in 2010 USAREC began assigning missions to the station level, a change from individual recruiter missions in prior years. Second, the time constraints of our research would not have permitted the use of 2015 data. That data would not have become available until well into our analysis phase. In most cases, USAREC was able to provide monthly observations for 2010–2014 which—when using time (t) as the observational unit—results in a sample size (T) of 60 for each unit. Therefore, we meet an important requirement for time series data since it is recommended that $T \geq 50$ [40].

Mission Variable Metrics. There are two major components of mission metrics: the contracts missioned and the contracts achieved. The mission expresses the goal for how many Enlistment contracts of each types are to be produced by each unit.[6] The "achieved" is the number of contracts actually attained. USAREC only assigns contract missions to brigades at annual intervals, but each brigade is free to adjust the missions of its subordinate echelons, and does so on a quarterly basis. We received monthly unit "adjusted" mission and achieved data for each of the three categories, resulting in the group of six mission metrics from Table 2.

Enemy (Competition) Variable Metrics. Previous literature has included metrics which describe competition for the youth market from Sister-service recruiters and even civilian employers. USAREC obtained for us—from DMDC—annual data for the numbers of contracts and recruiters in each battalion area, by Service and component. The current USAREC model indirectly accounts for past performance of a market with respect to the Army's share of contracts vs. other Services. We define contract share (z_{21}) generally as the number of Army contracts divided by the total number of contracts for all Services, with one main caveat: we include total AC contracts in the numerator and AC + RC contracts for *all* Services in the denominator, since AC contracts are in competition with RC contracts, in some sense. Also, we note importantly that the contract share data does not distinguish between education and aptitude categories. We define the percent recruiter share (z_{22}) as the number of Active + Reserve Army Recruiters divided by the total Active + Reserve recruiters from all Services; we do not separate the Active and Reserve Army components in this case since a station with both components shares a common mission.

[6] A contract is not the same as an Enlistment. Unless a contract signee ships immediately to basic training, he or she is placed in the Delayed Entry Program (DEP) and can decide not to enlist. This is known as a "DEP-loss," and is beyond the scope of our research. Suffice it to say that USAREC must set its contract mission above the required number of accessions (i.e., Enlistments) in order to account for DEP attrition.

Troops and Support Variable Metrics. We now address a few metrics related to the manpower and effort considerations of USAREC. Past studies found the number of Recruiters to be significant in affecting enlistments. Fortunately, USAREC was able to furnish its recruiter strength by month. Therefore, we define the metric "Army Recruiters" as the number of AC + RC recruiters on-hand (z_{23}). Two other metrics are also useful: appointments made and appointments conducted. Upon receipt of the recruiting mission, recruiters schedule and conduct face-to-face appointments with prospective Enlistees. If a prospective Enlistee wants to continue pursuing the Army after an appointment is conducted, then he or she undergoes a series of physical and other eligibility exams before an Enlistment contract can be executed. None of the latter steps can [doctrinally] take place without an appointment; consequently, USAREC leaders use appointments made and conducted as indicators of recruiter effort. These metrics were not included in previous literature we reviewed, but we include it based on previous findings regarding the significance of recruiter effort as well as input from several recruiting subject matter experts (SMEs).[7] We received this data at the station level and aggregated it to battalion and brigade echelons, respectively.

Time Variable Metrics. Time plays an inherent role as an independent variable in our research by virtue of its use as an observational unit in the data cross-section. However, we also include as our last independent variable one additional time-related metric—that of *processing days* (z_{26})— in the mission variable set because it has been included in prior versions of USAREC's current missioning model. This variable is defined as the number of days which are available to conduct administrative enlistment processing activities and can generally be thought of as simply

[7] In addition to personnel in the USAREC headquarters, we interviewed local recruiting personnel as well as one current and three former recruiting company commanders.

the number of work days, assuming no over-time. We include z_{26} at the monthly unit of observation while USAREC models it at the quarterly interval. In order to make the simple conversion, we calculated average work days per month, subtracting for appropriate extended weekends and Holidays per standard U.S. Government observances.

At this point, we have concluded our metric descriptions. Before moving on to a discussion of our specific mathematical techniques, however, we now present a brief overview of how we amassed and structured the various metrics just discussed.

Database Structure.

We imported, weighted, and imputed our data in a macro-enabled workbook file of Microsoft Excel®2010. According to the unique structures and large sizes of our datasets, we wrote several subroutines in VBA to automate data pre-processing. In order to stream-line error checking and guard against erroneous data entry we maintained a separate worksheet for each unit, appending each unit with new data as it was processed. This structure proved to be well-suited for extraction to JMP® statistical software. The pseudo-code for our data organization procedure is given below:

```
WITH a macro-enabled spreadsheet
   FOR EACH Battalion
      create worksheet
      FOR EACH Variable
         IF Variable data == county-level
            FOR EACH time unit IN Variable
               IF number of counties <> number of observations
                  impute missing values with hot deck method
               END IF
            NEXT time unit
            align to battalion-level using weighting scheme
            IF Variable <> monthly
               impute missing months with stochastic mean value method
            END IF
         ELSE
         END IF
         store Variable
      NEXT Variable
   NEXT Battalion
END WITH
```

We now depart from the data and move on with a discussion of the quantitative methods we applied. We begin with variance reduction techniques and continue with model estimation, variable selection, concluding with model validation.

3.4 Variance Reduction

Our need for variance reduction techniques arises from the likelihood—given our large number of prospective variables—that there will be correlation between independent variables. Hence, we will likely require a means of reducing this inter-dependency in order to generate adequate mathematical models; we find that the multivariate technique known as Principal components analysis (PCA) is effective to achieving this end. PCA extracts p weighted linear combinations—called *components*—of a set of p variables, such that (1) each component accounts for a successively smaller amount of the total variance in the original dataset when placed in decreasing order, and (2) the components are uncorrelated with (i.e., *orthogonal* to) each other. The functional relationship under PCA is expressed by

$$PC_{(m)} = w_{(m)1}X_1 + w_{(m)2}X_2 + \cdots + w_{(m)p}X_p \tag{3}$$

where $PC_{(m)}$ is the mth principal component and $X_1, X_2, \ldots, X_p$ are the original variables. The $w_{(m)j}$ are weights applied to each original variable to as to maximize the ratio of variance of PC_m to the total variation subject to $\sum_{j=1}^{p} w_{(m)j}^2 = 1$, for $m = 1, 2, \ldots, p$. In analyzing the variance of each ordered p component, we may reach a point where the cumulative variance of the dataset explained at component p^* is satisfactorily high even though $p^* < p$. In such a case, we could discard the remaining $p - p^*$ components without much loss in the original data's information. Therefore, let p^* be the number of *retained* principal components which is less than

the number of original variables while still accounting for a majority of the total variance in the original set of variables [41]. This property makes PCA a potentially useful data reduction technique. The components themselves are not designed for explicit interpretation other than to explain the majority and relative direction of variance in the original variables. However, occasional interpretation is possible with component loadings [42].

To describe component loadings, let $\mathbf{R}$ be the correlation matrix of independent variables $X_1, X_2, \ldots, X_p$. Then the loading of variable i on component j is given by

$$w_{ij} = a_{i(j)}\sqrt{l_{(j)}} \tag{4}$$

where $a_{i(j)}$ is the ith element of the eigenvector associated with component j and $l_{(j)}$ is the eigenvalue of component j. Loadings are useful inasmuch as their magnitudes indicate how much a particular variable is affecting the variance of each component relative to the other variables. Component scores, denoted by $\mathbf{Y}$, orient the component loadings to new orthogonal axes. If obtained from the correlation matrix, the component scores of the r retained components are given by

$$\mathbf{Y} = \mathbf{A}\mathbf{X}_\mathbf{S} \tag{5}$$

where $\mathbf{A}$ is the matrix of eigenvectors of the p^* retained components and $\mathbf{X}_\mathbf{S}$ is the standardized data matrix of the same [41]. There are several methods to determine p^*. We use Horn's criteria, a continuous curve created by sampling the eigenvalues from K sets of normally and independently distributed (NID) random variates of dimension equal to the original dataset, whose correlation structure is characterized by an identity matrix [41]. We choose $K = 1000$ by convention [42].

3.5 Model Estimation

Ordinary Least Squares.

Our first objective is to find an adequate mathematical model that describes the effects of a battalion's market conditions on contract production. We also want this model to accurately predict future values of contract production. A multiple linear regression model of the market factors $x_1, x_2, \ldots, x_k$ on contract production, y, takes the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{6}$$

where ε is an error term that is normally and independently distributed with a mean of zero and constant variance (NID,$0, \sigma^2$) [43]. The values of $\hat{\beta}_j$, for $j = 1, 2, \ldots, k$ are the estimated *regression coefficients*, which express the per unit change in y for the corresponding x_j when all other regressors (i.e., $\forall x \neq x_j$) are held constant. OLS estimates the values of the regression coefficients so as to minimize the sum of the squares of the differences between each actual (y) and predicted ($\hat{y}$) value pair. This is equivalent to minimizing the sum of squared errors since $(y - \hat{y})^2 = \varepsilon^2$ [43]. When the model is given in its estimated form, we use $\hat{y}$ in place of y and omit ε since the prediction of a new observation in an adequate model is only dependent on the estimated regression coefficients.

To obtain the regression coefficient estimates ($\hat{\beta}_j$), we note that (6) may be re-written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{7}$$

from which it can be shown that the least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{8}$$

assuming that $\mathbf{X}$ is a $N \times p$ matrix with full column rank (i.e., p independent columns). The form of (8) is known collectively as the system of least squares (LS) equations. We have introduced matrix notation since it will be helpful in later discussions of model adequacy, where we note special properties of elements within the least squares equations. We discuss shortly the method for evaluating the output of the LS equations. First we pause for a few brief excursions into alternate forms of OLS for second-order response models, categorical variables, coded units, and centering.

The form of (6) is known as a first-order regression model without interaction. If non-linearity is detected in the fit of particular regressors, we may correct this issue by developing a second-order response model of the form

$$y = \beta_0 + \sum_j \beta_j x_j + \sum_j \beta_{jj} x_j^2 + \sum_{i \neq j} \beta_{ij} x_i x_j + \varepsilon \tag{9}$$

where the second and third summations add second-order quadratic and first-order interactions, respectively [44].

We estimate the models given by (6) and (9) for a sample size of N observations using both continuous and categorical variables. Categorical variables can be modeled together in a single model or, equivalently, by separate models for each category. An advantage to the former option is the ability to interpret a single set of summary statistics describing the total fit adequacy across all regions, but in our research we use both formulations. We now illustrate the use of categorical and continuous factors within the context of our research by assigning $n - 1$ indicator variables to denote n recruiting battalions (categorical variables). An indicator variable x_j is binary, with one of the n categories serving as the "baseline." A notional example of this is shown in Table 7 where Battalion A is the baseline and Battalions B and C are two other categorical assignments.

Table 7 begins with x_2 because we have also included x_1 as a hypothetical inde-

Table 7. Example of Indicator Variables

	x_2	x_3
Battalion A	0	0
Battalion B	1	0
Battalion C	0	1

pendent variable that is continuous and quantitative (applicable to all units), in order to subsequently illustrate how the two types of variables interact. In the current construct, the basic model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$. This is actually also the model for Battalion A since its intercept term is no different than $\hat{\beta}_0$. This is easier to see if we now add the indicator for Battalion B:

$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\ \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 (1) \\ \hat{y} &= (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 x_1 \end{aligned} \tag{10}$$

In (10), we assume that the slopes of each battalion are equal. However, we can add additional complexity—perhaps fidelity—by also allowing the slopes of the continuous coefficients to change. This is accomplished by allowing the indicator to interact with each continuous term. Consider the following scenario for Battalion C:

$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3 + \hat{\beta}_{13} x_1 x_3 \\ \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 (1) + \hat{\beta}_{13} x_1 (1) \\ \hat{y} &= (\hat{\beta}_0 + \hat{\beta}_3) + (\hat{\beta}_1 + \hat{\beta}_{13}) x_1 \end{aligned} \tag{11}$$

Now regardless of whether indicator variables are present, it is important to note that the LS estimates β_j are interpreted as the estimated change in y per unit change in x_j, given that x_j is in its natural units. While clearly useful from a practical standpoint, the use of "natural" units renders conclusions about the relative importance of

one independent variable to another impossible given different units of measure. We can resolve this problem by standardizing all x_j on a $(-1, 1)$ scale. By letting ξ_j be the value of the variable j in natural units, we have the following conversion to coded units, x_j:

$$x_{j,t} = \frac{\xi_{j,t} - [(\xi_{jMax} + \xi_{jMin})/2]}{[(\xi_{jMax} - \xi_{jMin})/2]}, \quad \text{for } t = 1, 2, \ldots, N \tag{12}$$

In this case, we use $\hat{b}_j$ to denote the LS estimates using coded variables. With scaling equalized between all variables, the magnitudes of each $\hat{b}_j$ can be assessed for relative importance [44]. We must be cautious in our interpretations, however, as the coefficients are only as good as the sample data from which they are obtained and may not be valid over the entire range of the independent variables [43].

As a final note on forms of the OLS model, we address the technique of centering. Centering can be applied to continuous variables in either natural or coded units, and is often necessary to reduce interdependency caused by ill-conditioning of the matrix $\mathbf{X}$. Such ill-conditioning is common with polynomial or interaction terms, and is implemented by replacing the observation x_t with an adjustment for its overall mean, or $(x_t - \bar{x})$. Centering is less interpretable but reduces variance inflation of the independent variables [43].

Hypothesis Testing.

The model in (6) and (7) is useful if there exists a linear relationship between the response, y, and the regressors. Thus, we apply a test of statistical significance to determine if such a relationship may indeed exist. In this test, we evaluate the null hypothesis (H_0) that all regression coefficients are zero against the alternate hypothesis (H_1) that at least one regression coefficient is different from zero. This

hypothesis test is expressed by [43]

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 &: \beta_j \neq 0 \quad \text{for at least one } j \end{aligned} \tag{13}$$

The residual sum of squares, SS_{Res}, which was minimized in (6), accounts for the error in LS estimation. Hence, we also have a regression sum of squares, SS_R, which accounts for the changes in the response being captured by the model. It can also be shown that each of these terms, when divided by its respective variance (σ^2), follows a χ^2-distribution with k and $n-k-1$ degrees of freedom for SS_R and SS_{Res}, respectively. Then, by definition of the F-statistic, we have

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} \tag{14}$$

which follows the $F_{k,n-k-1}$ distribution under a true null hypothesis [43]. Let α be the probability of a Type I error, defined as a rejection of the null hypothesis when the null hypothesis is true. Unless otherwise stated, we set $\alpha = 0.05$ by convention. We can then evaluate (13) by comparing the value of F_0 with $F_{\alpha,k,n-k-1}$. We reject H_0 when $F_0 > F_{\alpha,k,n-k-1}$, and fail to reject otherwise [43].

In the event that H_0 from (13) is rejected, we must subsequently determine which β_j terms are of real value in affecting the response. Each regressor in the model increases the variance of the predicted response ($\hat{y}$), which we would ideally like to minimize while still achieving predictive capability. Thus, we employ the following test for any regression coefficient:

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0 \end{aligned} \tag{15}$$

The test statistic used to evaluate (15) follows a t-distribution with $t \sim t_{\alpha/2,n-k-1}$ given a true H_0, and is given by

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \tag{16}$$

where C_{jj} is the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to $\hat{\beta}_j$ and the denominator is the *standard error* of the regression coefficient $\hat{\beta}_j$. The test is rejected when $|t_0| > t_{\alpha/2,n-k-1}$. A rejection of H_0 indicates sufficient statistical evidence—based on sample data—to conclude that the regressor x_j is explaining part of the variation in y. A failure to reject H_0 indicates insufficient statistical evidence to conclude x_j is having an effect on y; this term should be considered for elimination from the model as it is increasing the variance in the predicted response without adding new information [43].

In our analysis we use JMP®11 statistical software, which reports P-values for all hypothesis tests. A P-value indicates the smallest value of α for which the null hypothesis should be rejected [30]. Thus, it adds information by showing the magnitude of evidence in support of a conclusion regarding statistical significance. For this reason, we also report P-values for hypothesis tests in our analysis.

Model Adequacy.

In this section, we briefly address a few important aspects concerning model adequacy. By adequacy we mean several things, all of which must be addressed for the model to have value as a stable and reliable analysis platform:

1. General assessment of the model's fit to the data used in its estimation (or the model's fit to a separate *validation* dataset as discussed in Section 3.7)
2. Conformity to the major assumptions of linearity and the residual terms having NID $\sim (0, \sigma^2)$ structure [43]

3. Analysis of diagnostics for leverage and influence of individual observations

4. Reduction of multi-collinearity between the regressors

We now discuss our methodology for dealing with each adequacy consideration, in kind.

General Assessment Metrics. The coefficient of determination, R^2, is the ratio of SS_R to SS_T. Thus, it is the percent of the total variation in the data which is explained by the specified regression model. However, it can be shown that R^2 is artificially inflated by adding non-valuable terms to the model. Therefore we select an alternative metric for general fit assessment, the adjusted R^2:

$$R^2_{Adj} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2) \tag{17}$$

where p is the number of independent terms in the model, including the intercept [43]. This metric is more appropriate for our use since it accounts negatively for the addition of extraneous terms to the model, and we have a potentially large p with all possible indicator-continuous variable combinations considered.

We also make use of the "R^2-like" statistic known as Prediction R^2. This metric is suitable for our use given that time is our observational unit and that the chief purpose of our model is to make predictions about future observations. It is defined as

$$\begin{aligned} R^2_{Pred} &= 1 - \frac{PRESS}{SS_T} \\ &= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{(i)})^2}{SS_T} \end{aligned} \tag{18}$$

where $PRESS$ is the prediction sum of squares. The value of $PRESS$ is calculated from the residuals when observation i is predicted without its use in the dataset

(denoted by the subscript in parenthesis). Thus, it is a form of data-splitting which aids in assessment of predictive performance [43]. If the model is a stable predictor, we expect the values of R^2_{Adj} and R^2_{Pred} to be relatively close to one another. Of course, this expectation is predicated by an assumption that the underlying process in the data remains constant between periods. Otherwise, large deviations in both metrics may be observed.

Residual Analysis. Residual analysis consists mainly of verifying three assumptions about the residuals ($e_i = y_i - \hat{y}_i$) which relate to normality, independence, and constancy of variance. Deviations from these assumptions could cause the model to be unstable and incorrectly estimate the parameters. Therefore, we devote considerable attention to ensuring conformity.

First, the residuals must be normally distributed with a mean of zero. We perform this check by visual inspection of a histogram and a normal probability plot of the internally studentized residuals, r_i. The advantage of using r_i as opposed to the "raw" e_i residuals is that the former accounts for the distance of each observation from the centroid of the independent variable data cloud. Thus, the r_i are less susceptible to small residual variances that results when remote points pull the regression equation to themselves [43]. The definition of r_i is given by

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1 - h_{ii})}} \tag{19}$$

where $MS_{Res} = SS_{Res}/n - k - 1$ and h_{ii} are the diagonal elements of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. It can be shown that h_{ii} gives an expression for relative distance of the ith observation from the center of the data [43]. The additional ease with which the r_i are computed in JMP® makes them a logical choice for residual analysis [29].

Constancy of variance is initially checked by visual inspection of the predicted

values, $\hat{y}_i$, plotted vs. the r_i. Violations of this assumption are indicated by irregular patterns in the plot, such as a funnel, double-bow, or non-linearity. A violation of this assumption likely requires a corrective transformation on y and/or the regressors to stabilize the variance. We use the 95% confidence interval on λ, the parameter in Box-Cox transformation methods where $y' = y^{\lambda}$ [43]. While we clearly need a selected transformation to stabilize the variance, we also require that it be palatable to decision-makers. Using a confidence interval on the value of λ gives us flexibility to choose, say, $y' = y^{1/2} = \sqrt{y}$ even though the point estimate for λ may be 0.4.

Independence of the residuals is a primary concern in our analysis since we use time as the observational unit. We use the Durbin-Watson test to check for the presence of first-order autocorrelation, defined as correlation between errors that are one time period apart. A first-order autoregressive process is defined as

$$\varepsilon_t = \phi\varepsilon_{t-1} + a_t \tag{20}$$

where ε_t is the residual obtained from OLS regression in time period t, ϕ is an autocorrelation parameter that must be estimated with OLS, and a_t is a NID$(0, \sigma^2)$ random variable. The estimates of error variance and root mean square error in (20) are $\hat{\sigma}_a^2$ and $\hat{\sigma}_a$, respectively; we make use of the latter in our discussion of forecasting prediction intervals [29]. The Durbin-Watson test evaluates the following hypotheses:

$$\begin{aligned} H_0 &: \phi = 0 \\ H_1 &: \phi \neq 0 \end{aligned} \tag{21}$$

Since we estimate the two-tailed alternate hypothesis, the Type I error is 2α; we choose $\alpha = 0.025$ to obtain an overall 0.05 Type I error probability. The Durbin-Watson statistic, d, is used to evaluate (21) for positive autocorrelation; $4 - d$ can be

simultaneously used for negative autocorrelation with

$$d = \frac{\sum_{t=2}^{N}(e_t - e_{t-1})^2}{\sum_{t=1}^{N} e_t^2} \tag{22}$$

where N is the total number of observations [29]. The value of d depends on the $\mathbf{X}$ matrix, but can generally be shown to lie between lower- (d_L) and upper- (d_U) bounds depending on values of α, $p-1$, and N. We use $p-1$ to denote the number of regressors, which is the number of parameters less the intercept. The hypothesis test then proceeds as follows:

$$\text{If } (d \text{ or } 4-d) < d_L, \text{ reject } H_0$$

$$\text{If } (d \text{ or } 4-d) > d_U, \text{ do not reject } H_0$$

$$\text{If } d_L \leq (d \text{ or } 4-d) \leq d_L, \text{ the test is inconclusive}$$

In the event that first-order autocorrelation is present, we add the lag 1 value of the response (y_{t-1}) as an additional regressor in the model. This is the recommended strategy before much more complex forecasting techniques become necessary [43].

Leverage and Influence. Both leverage and influential points are observations that have unusual values in x-space. Leverage points do not affect regression coefficient estimates but do affect the coefficient standard errors, as well as model summary statistics like R^2. We define a leverage point as any observation for which $h_{ii} > 2p/N$, or twice the average of the diagonal of $\mathbf{H}$. By contrast, influential points do affect the regression coefficients because they "pull" the regression model in their direction. We use Cook's $D > 0.25$ as the criteria for determining influence. The definition of Cook's D for an observation i is

$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}, \quad i = 1, 2, \ldots, n \tag{23}$$

It is possible that both leverage and influence points are the result of incorrect data entry or collection. However, it is also possible that they are legitimate and warrant further analysis. We have chosen not to eliminate any observations from our model, given our limited knowledge of USAREC's data entry/collection process. Therefore, our analysis of leverage and influence points is strictly informational, in that we identify those points warranting further investigation by subject-matter experts.

Multicollinearity. Multicollinearity is the presence of linear or near-linear dependencies between regressors. When multicollinearity is absent, the regressors are orthogonal. That is, they are perpendicular to each other in a multi-dimensional sense. Multicollinearity must be diagnosed and corrected to the greatest extent possible. Otherwise, estimates of the regression coefficients and their (co-)variances could be seriously inaccurate. One interesting aspect of the multicollinearity problem is that it can often be disguised by a seemingly excellent summary statistic (i.e., an R^2 close to unity). The presence of multicollinearity alone does not mean that the model will be a poor predictor, but this is often the case [43].

Given our initial data collection of nearly two dozen variables, we expect some multicollinearity in the dataset. We might also expect multicollinearity in the event that polynomial terms are added to a model and are not centered. However, we can also examine variance inflation factors (VIFs) for each regressor. A VIF is defined for each j regressor as

$$VIF_j = C_{jj} = (1 - R_j^2)^{-1}, \quad j = 1, 2, \ldots, k \tag{24}$$

where C_{jj} are the diagonal elements of the covariance matrix $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$, and R_j^2 is obtained by regressing x_j on the remaining $p - 1$ regressors. VIFs that are greater than 10 are generally cause for some corrective action, and there are a few methods for

dealing with high VIFs. Some methods, such as further data collection, are not feasible in our case given project time constraints. However, model re-specification is one method which we are able to apply. We employ two types of model re-specification: redefining and eliminating variables [43]. The variance reduction techniques given in Section 3.4 assist in both of these efforts since they reveal the magnitude and direction of the primary drivers of variance in the dataset. For example, PCA can suggest a reduced variable set by choosing one variable from each principal component, or by combining several variables from one component into a single index. In this way, the problem of multicollinearity is also one of proper variable selection. This provides a fitting transition to our discussion of mixed stepwise regression.

3.6 Variable Selection

Thus far in our treatment of basic OLS estimation tenets, we have assumed that the variables in the model are all thought to be important. However, we have a large pool of possible regressors. As mentioned in our discussion of multicollinearity, some of these regressors are likely producing effects which can be effectively duplicated by other regressors. Therefore, we require some judgment as to what the *best* set of regressors is, inasmuch as it (1) produces the most stable model with highest predictive capability and (2) contains as few regressors as possible to reduce the variance of $\hat{y}$. What we have just described is the variable selection problem [43].

We choose the stepwise method as our vehicle for variable selection. Stepwise selection is one of a few variable selection procedures in which, at each "step," all candidate variables are assessed for their values of t_0, where t_0 is the model-fitting criteria. The analyst selects critical values of t by setting α and these values are used as entry and/or exit criteria for each t_0. Mixed stepwise regression is a modification of forward selection. In forward selection, variables are entered in order of highest degree

to which $|t_0| < t_{\text{IN}}$ where t_{IN} is set by the analyst. Since the forward-selected model could grow to be quite large, mixed stepwise regression requires an extra step. This modification requires variables previously entered into the model be subsequently re-assessed at each step; if a currently used variable is found to have $|t_0| > t_{\text{OUT}}$, this variable is discarded. The procedure continues until no variables can be added or discarded. In our application, we set $\alpha_{\text{IN}} = 0.05$ by convention and $\alpha_{\text{OUT}} = 0.1$, making it relatively difficult for inclusions while allowing some leniency prior to exclusion. This is often recommended as it reflects added emphasis on model parsimony [43]. We use previously discussed fit metrics R^2_{Adj}, R^2_{Pred}, and add Mallow's C_p which is given by

$$C_p = \frac{SS_{Res}(p)}{\hat{\sigma}^2} - T + 2p \tag{25}$$

to describe the model fit by stepwise selection. We include C_p for its simple interpretation. It can be shown that desirable values of C_p are small (i.e., in the vicinity of or less than p) [43].

We have chosen stepwise selection for several reasons. First, it is less computationally demanding given the size of our dataset than an all possible models approach. Second, it combines the best elements of forward selection and backward selection procedures (backward elimination begins with all regressors in and eliminates them based on t_{OUT} [43]. Third, stepwise selection is convenient since we already have baseline models developed by USAREC. We are able to manually enter the existing USAREC variables and then let stepwise selection take over with our augmented variable set. Finally, there is not universal agreement among experts over the best procedure and none is guaranteed to produce a truly "best" subset of regressors [43]. However, we can use the unique structure of our data to our advantage regarding this issue. Since the USAREC data is divided by 38 mutually exclusive regions, we may attempt to fit separate models for each region. Comprehensive analysis of the selected

models together may help reveal the prominence of certain selected regressors and, consequently, a universally acceptable subset.

At this point, we have covered the relevant aspects of the model building process. However, even the best built models—if they cannot achieve their primary purpose of accurately describing a process—may not useful. Therefore, we now transition to the final portion of our methodology with a discussion of model validation techniques.

3.7 Model Validation

Data Splitting. The primary purposes of our models are to predict future data. In that sense, validation can be described as how well the estimated model performs in the presence of "future data." Since we are limited by an inability to augment the database with new observations in real time, we elect to split the existing data. In data splitting of time series data, we let observations $t = 1, 2, \ldots, T$ define the estimation set. This set is used in the model building processes described in previous sections. The remaining observations $t = T+1, T+2, \ldots, T+\tau$ define the validation set. The validation set has no part in estimating model parameters or selecting variables; it is strictly used to "test" the performance of the model gained from the estimation set [43]. In our dataset, we let $T = 45$ and $\tau = 15$ define the estimation and validation sets, respectively. This split is not arbitrary for two reasons. First, at least 15 to 20 observations are recommended to gain an adequate assessment of prediction performance [43]. Second, USAREC begins setting missions a few months prior to the next full recruiting year. By adding three months to the validation set, we effectively re-create the decision situation from the headquarters point of view: USAREC must attempt to predict contract production over an extended planning horizon using only the data realized by the decision-point, T.

Validation Metrics. Because our models use time series data, "new" data inherently means "future" data. Making a prediction about future events is a *forecast* and the usual metrics of model fit such as R^2_{Adj} and R^2_{Pred} do not apply since forecast data are not used to fit the model itself [29]. For this reason, we now introduce several metrics which are useful specifically for assessing forecasting accuracy. We begin with two metrics which are scale-dependent; that is, their interpretation depends on the units in which the forecast is made. The first of these is the Mean Absolute Deviation (MAD), defined by (26)

$$\text{MAD} = \frac{1}{N} \sum_{t=T+1}^{T+\tau} |y_t - \hat{y}_t| \tag{26}$$

where y_t is the actual response at lead time $T+1, T+2, \ldots, T+\tau$ from origin time T and $\hat{y}_t$ is the predicted value of the same [45]. While it does present a measure of central tendency for the forecast errors, the MAD lacks information regarding the spread, or variability of the forecast errors. Therefore, we also use the forecast Mean Square Error (MSE) and Root Mean Squared Error (RMSE) as measures of variability in the errors. The RMSE has the added advantage that is interpreted as the standard deviation of forecast errors (not squared units as with the MSE) [29].

$$\begin{aligned} \text{MSE} &= \frac{1}{N} \sum_{t=T+1}^{T+\tau} (y_t - \hat{y}_t)^2 \\ \text{RMSE} &= \sqrt{\text{MSE}} \end{aligned} \tag{27}$$

The MAD and RMSE are useful metrics insofar as they provide contextual understanding of forecast errors in the same units of the forecasts. However, in the absence of scaling by the actual value (y_t), neither metric provides an understanding of the *magnitude* of forecast error. Thus, it is not possible to compare forecasts between differing categories or time periods using MAD or (R)MSE. Hence, we also use the Mean Absolute Percent Error (MAPE). MAPE describes the average accuracy of a

particular forecast over a period of time, relative to the observed data. The definition of MAPE is [29]

$$\text{MAPE} = 100\% \cdot \frac{1}{N} \sum_{t=T+1}^{T+\tau} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \tag{28}$$

where previous definitions of each variable still apply. Thus, we use MAPE to more effectively compare forecast errors across differing categories (e.g., recruiting regions and contract types). Finally, we also find the coefficient of determination, R^2 to be useful for at least two reasons. One is its ubiquity throughout regression literature, thus making it useful for comparisons of our results to both past and future research. The second reason concerns its relative comparability to the fit metrics of the estimation data, R^2_{Adj} and R^2_{Pred}. A relatively good forecast should have R_2 values which are generally consistent with—even if slightly lower than—the metrics of fit used in the estimation data. Earlier we defined $R^2 = SS_R/SS_T$ but note that this can be expressed alternately as [45]

$$\begin{aligned} R^2 &= SS_R/SS_T \\ &= 1 - \frac{SS_{Res}}{SS_T} \\ &= 1 - \frac{\sum_{t=T+1}^{T+\tau} (y_t - \hat{y}_t)^2}{\sum_{t=T+1}^{T+\tau} (y_t - \bar{y})^2} \end{aligned} \tag{29}$$

As a final word of note regarding (26-28), the number of observations N can be defined in multiple ways when categorical variables are involved, depending on the level of desired analysis. Thus, when analyzing errors over the entire dataset, $N = \tau B$ where τ is the number of time periods and B is the number of categories (e.g., recruiting battalions). But when forecast errors are evaluated for individual recruiting battalions, the number of observations is simply $N = \tau$. Note that in

all cases, we sum over only the time periods in the validation data set. Thus, the predictions made in (26-28) are separate from the predictions obtained during the model fitting process.

Prediction Intervals. While MAD, MAPE, etc. are adept at describing the fit of point predictions to the data, we also would like to provide an indication of the uncertainty surrounding any given point estimate. Prediction Intervals (PIs) are appropriate for this task since they specify a probability that a realized future value will lie between upper and lower bounds. There are competing methods for forming prediction intervals. What often differentiates these methods are two considerations; the first of these is the lead time (i.e., the number of periods in the future for which the forecast is made). We assume a lead time of 1 month since USAREC adjusts its missions on a monthly basis.[8]. The second consideration involves whether or not the inputs are known at the time the prediction is made. If one assumes the inputs to be known, then the only error present in the prediction of future observations is due to the model's mis-specification. For example, for a one-period ahead forecast ($t = t+1$) with known inputs, the prediction interval is given by

$$100(1-\alpha)\%\ \text{PI} = \hat{y}_{t+1} \pm z_{\alpha/2}\hat{\sigma}_a \quad t = T, T+1, T+2, \ldots, T+\tau-1 \qquad (30)$$

where $\hat{\sigma}_a$ is the square root of the NID$(0, \sigma)$ mean squared error obtained by regressing ε_{t-1} on ε_t [29].

In our case, (30) is reasonable given monthly forecasts. However, it does not account for data unknown at the time of the forecast. Let us assume for a moment that in any given time period the factor x affects the number of enlistment contracts.

[8]Actually, USAREC headquarters only adjusts quarterly missions, but subordinate headquarters are not bound by this constraint and do adjust missions on a monthly basis. We make this assumption to reflect the additional flexibility given to subordinate echelons

But for a predicted number of contracts in time period $T+1$, the value of x_{T+1} would have to also be forecast if the prediction for contracts is made in the prior month at time period T. In such a case, the model now has error from its predictions of both the response *and* the independent variable. If we assume model and independent variable forecast errors to be independent for a one period-ahead forecast, we have a $100\%(1-\alpha)$ PI given by

$$100(1-\alpha)\%\text{ PI} = \hat{y}_{t+1} \pm z_{\alpha/2}[\hat{\sigma}_a^2 + \hat{\beta}^2\hat{\sigma}_x^2]^{1/2} \quad t = T, T+1, T+2, \ldots, T+\tau-1 \quad (31)$$

where aforementioned definitions apply for $z_{\alpha/2}$ and $\hat{\sigma}_a$; $\hat{\beta}^2$ is the square of the coefficient for x obtained from the original OLS model and $\hat{\sigma}_x^2$ is the NID$(0, \sigma^2)$ estimate of error variance obtained from the first-order autoregressive model of x [29]. However, the citation provides (31) based on only one forecast independent variable. We therefore generalize the form of (31) for multiple regressors x_j in the set $j = 1, 2, \ldots, m$ regressors which must be forecast:

Proof. A well-know property of the variance, V, of a random variable X is that $V[aX + b] = a^2V[X]$ where a and b are constants [46]. Additionally, the variance of a sum of independent random variables $X_1, X_2, \ldots, X_m$ is simply the sum of their individual variances such that [46]

$$V[X_1 + X_2 + \ldots + X_m] = V[X_1] + V[X_2] + \cdots + V[X_m].$$

We let ϵ denote the random variable associated with forecast error due specifically to the model form $y = \hat{\beta}_0 + \sum \hat{\beta}_j x_j + \epsilon$, and let $\hat{\beta}_j X_j$ be the random variable associated with each $j = 1, 2, \ldots, m$ estimated coefficient-forecast regressor combination. Then, assuming independence between the error of each regressor forecast as well as the model error assuming known inputs, we have the variance of total forecast error

given by

$$V\left[\epsilon + \sum_{j=1}^{m} \hat{\beta}_j X_j\right] = V[\epsilon] + V[\hat{\beta}_1 X_1] + V[\hat{\beta}_2 X_2] + \cdots + V[\hat{\beta}_m X_m]$$

Note that $V[\hat{\beta}_0] = 0$ since $\hat{\beta}_0$ is a constant. Using the aforementioned property of variance, then, we have

$$\begin{aligned} V\left[\epsilon + \sum_{j=1}^{m} \hat{\beta}_j X_j\right] &= V[\epsilon] + \hat{\beta}_1^2 V[X_1] + \hat{\beta}_2^2 V[X_2] + \cdots + \hat{\beta}_m^2 V[X_m] \\ &= \sigma_a^2 + \hat{\beta}_1^2 \sigma_{x_1}^2 + \hat{\beta}_2^2 \sigma_{x_2}^2 + \cdots + \hat{\beta}_m^2 \sigma_{x_m}^2 \end{aligned}$$

Denoting the set of forecast regressors with $\hat{\mathbf{x}}$ and replacing each population variance with its sample estimate, we can then state the generalized form of (31) as

$$100(1-\alpha)\%\text{PI} = \hat{y}_{t+1} \pm z_{\alpha/2} \left[\hat{\sigma}_a^2 + \sum_{x_j \in \hat{\mathbf{x}}} \hat{\beta}_j^2 \hat{\sigma}_{x_j}^2\right]^{1/2} \qquad t = T, T+1, T+2, \ldots, T+\tau-1 \tag{32}$$

□

To implement the use of (32), we propose a simplest-case method for forecasting the inputs by the rationale that if the simplest method can be shown to be effective, more complicated methods may not be necessary. Therefore, we use the estimation dataset to fit a simple linear regression for each continuous regressor, using time period as the independent variable. This is a form of smoothing; in fact, it is the simplest type and is commonly known as a simple trend model with form

$$x_t = TR + \varepsilon_t \tag{33}$$

where x_t is the response, $TR = \hat{\beta}_0 + \hat{\beta}_1 t$ is the linear trend component with time as

the regressor, and ε_t is the NID$(0, \sigma^2)$ error term [45]. With this form, the estimated trend line's extrapolation into the validation set will produce input variable forecasts (hence our use of x as the response variable). We assume a single set of forecasts made at the end of the estimation set for all time periods in the validation set to be suitable as a simplest case method for producing independent variable forecasts.

Now prior to making forecasts, we would ideally analyze all input variable models for adequacy to the standard assumptions. However, this would be considerably burdensome in our case since we have over three dozen categorical variables in addition to time series data. Therefore, we have assumed all battalion input trend models to be adequate, with single intercept and trend coefficients, respectively, in order to complete our analysis in a timely fashion. An extreme difference in the widths of the prediction bands between (30) and (32) may indicate a significant departure from this assumption, or that of the single forecast origin.

At this juncture, we have explicitly stated our methodology for the analysis and mathematical modeling of Army Recruiting battalion markets. We can now conduct the critical portions of the model building process with the following tools and techniques:

1. a sound, doctrinal framework from which we gathered pertinent data;
2. a means of reducing the correlation between the large number of variables resulting from (1);
3. iterative regression modeling that can help identify the best subsets of variables from (1) and (2);
4. procedures for evaluating the adequacy of models obtained from (3);
5. the ability to validate practical effectiveness of the models resulting from (4)

As one might observe from this brief summary, these methods are not used in isolation. Rather, these quantitative techniques work simultaneously and cumulatively to tell

a comprehensive story about the data. We now focus our efforts on chronicling this story in the next chapter.

IV. Results and Analysis

4.1 Outline

In this chapter, we present the results obtained from applying the methodologies described in Chapter III. Our flow through the material is guided by a logical order of events, which corresponds closely to the order in which we performed the analysis. Overall, our results can be divided into two over-arching parts in which Part I consists largely of exploratory analysis concerning the appropriate selection of responses and regressors; Part II then covers the refinement and validation of the regression models containing the variables obtained in Part I.

Therefore, we begin Part I in Section 4.2 with an investigative analysis of the current regression model in use at HQ USAREC (i.e., the RMI). We begin here because the RMI has not been formally analyzed in literature, and also because we desire a baseline for quantitative comparison of our own models. In Section 4.3 we leave the RMI to discuss initial steps of our original analysis. We start with PCA on all dependent variable candidates and select the three contract types as our responses. We then implement an iteration of mixed stepwise regression for each battalion on each of the three responses. We conclude this section with a discussion of multicollinearity found among the independent variables during stepwise regression. Section 4.4 covers our attempts to remedy the multicollinearity of the regressor set through the application of PCA; in this instance we use PCA to redefine and reduce the set of regressors in order to achieve greater orthogonality and parsimony. At this point we propose a redefinition and reduction of the regressor set in order to increase their orthogonality. We conclude this section and Part I of our analysis by conducting a second iteration of mixed stepwise regression for a second-order response surface model containing the refined set of regressors. This iteration is also completed at the

battalion level for each contract type.

We begin Part II of our analysis in Section 4.5, where we introduce indicator variables as a means of specifying separate battalion models. We then conduct a final iteration of mixed stepwise regression for each contract type, including first-order terms from our reduced variable set and interactions of these terms with battalion indicator variables. The result is a single model (per response) that can be assessed for adequacy and fit, while still allowing for the derivation of individual unit models as discussed in Chapter III. Next, we address the required conditions for model adequacy; we find that transformations of all three responses and the inclusion of lagged responses are necessary to satisfy conditions of homoskedasticity and independence. We locate a few potential leverage points but do not find sufficient reason for their exclusion from the models.

Finally, in Section 4.6 we validate each of the three adequate models estimated over the course of previous sections. We introduce 15 months of data unused during the estimation process to test the predictive accuracy of each model, given assumptions both of known and unknown inputs. For unknown inputs, we find that a simple linear trend model of each input provides a prediction interval which is close in proximity to that of known inputs. We conclude this section and the chapter by providing a few observations on validation performance of individual battalion markets with respect to each contract type.

As a final word of note: in many of our scatter plots we utilize USAREC's five-tone color scheme to differentiate datapoints in each of the five recruiting brigades. Initially a matter of curiosity to us, this technique proved helpful regarding interpretations of the data at several junctures. Figure 6 provides the color scheme used along with the RSIDs of each battalion (a full list of units and headquarters locations is in Appendix A).

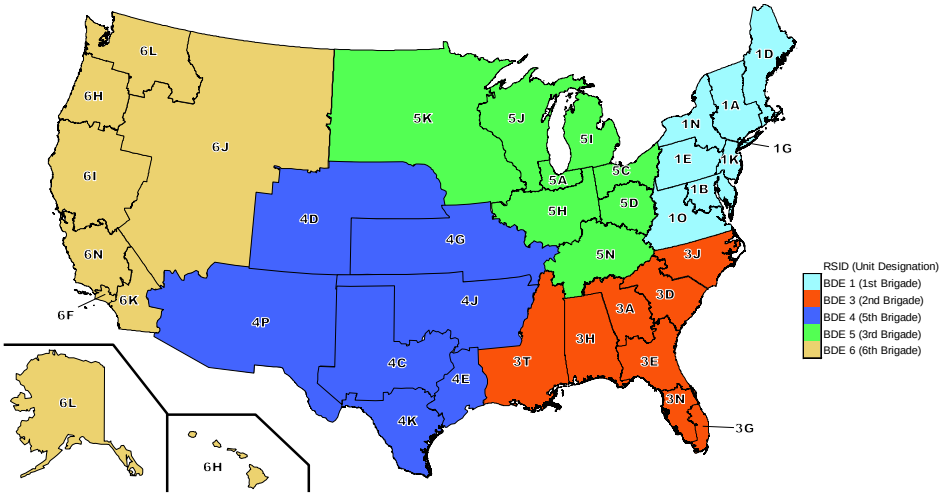


Figure 6. USAREC Boundaries as of the 1st Quarter, Recruiting Year (RY)2015

4.2 RMI Baseline

To establish a baseline for comparing our models, we fit a first-order linear regression model to the quarterly data provided to us by USAREC. We utilized the form of the RMI model specified in USAREC's documentation: the response is GA+SA contracts achieved per recruiter at the battalion-level; the independent variables are unemployment rate, propensity, battalion RSID, contracts required per recruiter (i.e., mission per recruiter), GA+SA contracts required per recruiter, and total contracts required per recruiter [4]. The first three independent variables correspond explicitly to x_4, x_{10}, and x_{13} from Figure 2, respectively. Using the notation from Figure 2, we define the other variables as follows for all units i and quarters t (scripts i and t are omitted for brevity):

$$\begin{aligned}
y_1 &\equiv \frac{z_{18} + z_{19}}{z_{23}} = \text{GA+SA contracts achieved per recruiter (GSA_PR)} \\
x_{27} &\equiv \frac{z_{15} + z_{16} + z_{17}}{z_{23}} = \text{total mission per recruiter (Req_Vol_PR)} \\
x_{28} &\equiv \frac{z_{15} + z_{16}}{z_{23}} = \text{GA+SA mission per recruiter (Req_GSA_PR)} \\
x_{29} &\equiv \frac{z_{18} + z_{19} + z_{20}}{z_{23}} = \text{total contracts achieved per recruiter (Vol_PR)}
\end{aligned} \tag{34}$$

Now we can express the form of the current RMI as $y_1 = f(x_4, x_{10}, x_{13}, x_{27}, x_{28}, x_{29}) + \varepsilon$. However, this form of the model is problematic because it places x_{29} in an independent role when it is really dependent. Both x_{29} and y_1 are defined by ratios of contracts achieved, which by definition are realized *as a result of* the effects of the other factors during a time period. The RMI already includes quality and total missions as independent variables, which are appropriate since they are determined at the beginning of each time period. However, the outcome of contracts achieved cannot also be included as an input in the same time period; this value is not known at the outset and is likely dependent upon the independent variables. This can be easily seen from (34), where the definitions of y_1 and x_{29} are identical save one term in the numerator.

Our solution to this issue was to fit two RMI models, one excluding and one including x_{29} in the independent variable set. The results of these models are presented in Figures 7 and 8 for the full and and reduced models, respectively, as well as in Table 8. In part (a) of Figures 7 and 8 we provide the fit summaries and in parts (b-c) graphical outputs for a brief residual analysis. Table 8 contains the detailed estimates for significant regression parameters (*incodedunits*) at the 0.05 significance level. The number of observations ($N = 608$) is the number of battalions (38) multiplied by the number of quarters, $q = 16$.

From visual examination and from the fit summary, it is apparent that the full model (including x_{29}) provides a superior fit to the data. Additionally, both models appear to satisfy the conditions of normality and constant variance, although some argument could be made for transformations amidst slight outward-opening funnels in both part (c). In part (c) we also label values of $r_i > |2.5|$ as outliers, assuming slightly more than 1% of the data assuming normality. While there are only a handful of outlying residuals in either case, we do note a tendency for these points to occur

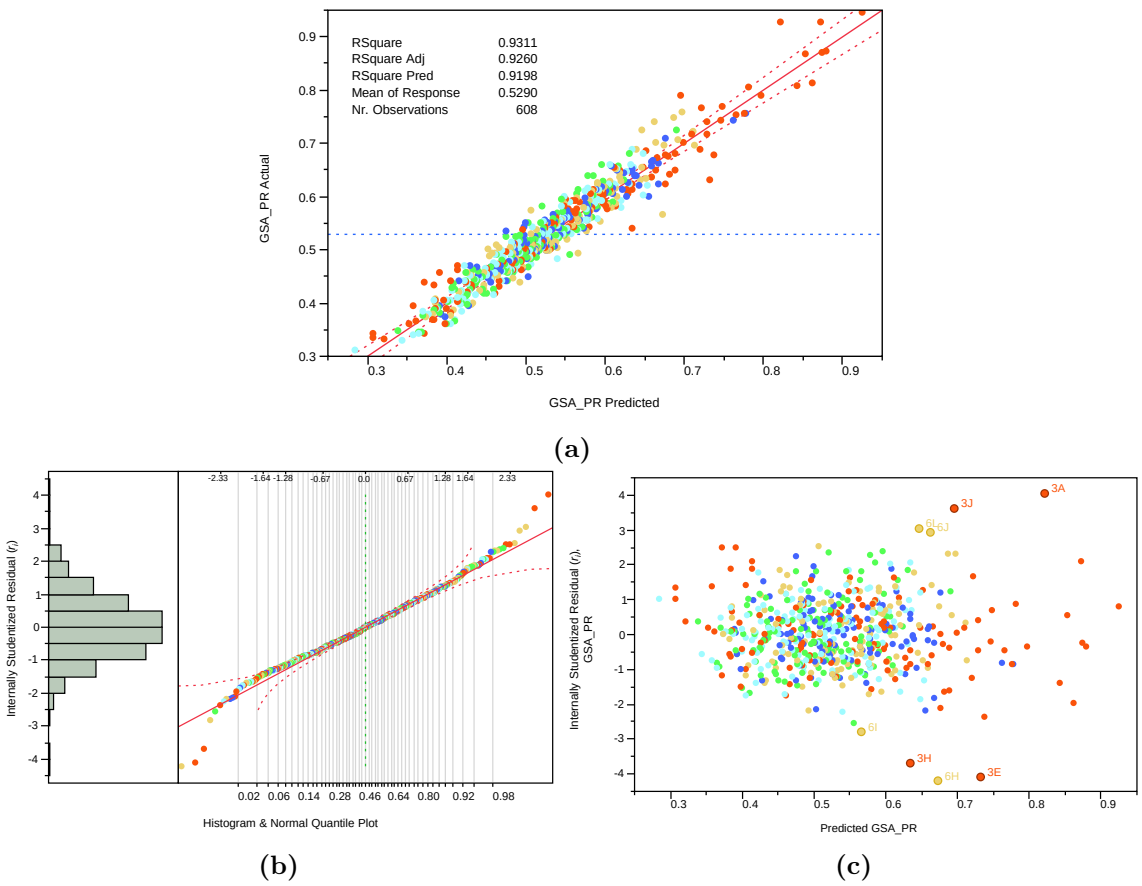


Figure 7. Fit Summary and Residuals (r_i) for the RMI, Vol_PR (x_{29}) Included

disproportionately in the battalions of 2nd BDE (e.g., BN 3A, 3H, 3J, etc.) and 6th BDE (e.g., 6H, 6I, etc.). Thus, there may be a unique effect in these particular units which is not being captured by either model.

In Table 8, we are primarily concerned with the signs and relative magnitudes of the $\hat{b}_j$'s, as well as the VIFs. We do confirm that both regressions are significant as well, since each has at least one $\hat{b}_j$ with a P-value < 0.05. The full RMI model is clearly dominated by the effect of x_{29}, which is twice the magnitude of the next most significant variable. Following 14 significant intercept shifts for units, the next two significant terms are the quality mission- and total mission-to-recruiter ratios. However, these terms are highly collinear as indicated by VIFs of 10.76 and 12.21, respectively. Therefore, this model could be mis-specifying the parameter esti-

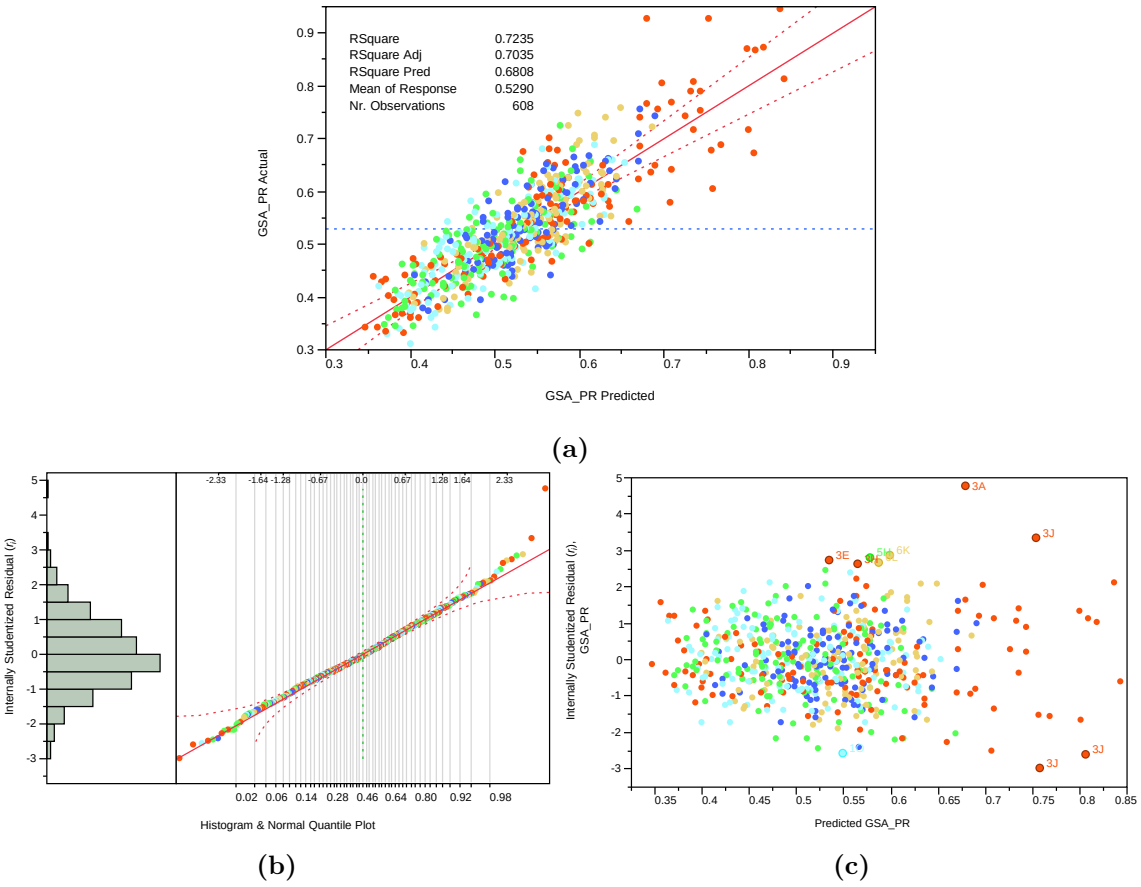


Figure 8. Fit Summary and Residuals (r_i) for the RMI, Vol_PR (x_{29}) Excluded

mates and under-estimating their standard errors. Regarding signs, we see a positive correlation between unemployment (x_4) and the response, which confirms intuition. The same is true for quality mission per recruiter, although in this model the total mission-to-recruiter ratio appears to have a negative effect on quality achieved per recruiter.

Now we examine the reduced RMI model parameters. When x_{29} is removed, unemployment (x_4) becomes the predominant main effect; this is obviously a direct reversal from the first model, although the positive sign is still appropriate. Other changes include the addition of propensity (x_{10}) as a significant effect and the apparent non-significance of quality mission per recruiter, which did appear in the full model. However, both propensity and total mission per recruiter terms have a positive effect

Table 8. Parameter Summaries in Coded Units for the RMI with (left) and without (right) x_{29}, Sorted in Decreasing Levels of Significance to $\alpha = 0.05$

Term (x_j)	$\hat{b}_j$	se($\hat{b}_j$)	P-value	VIF	Term (x_j)	$\hat{b}_j$	se($\hat{b}_j$)	P-value	VIF
BN_5N (Int.)	0.6386	0.0073	<.0001		BN_5N (Int.)	0.5738	0.0142	<.0001	
x_{29}	0.2949	0.0071	<.0001	4.0393	x_4	0.1604	0.0121	<.0001	3.7045
BN_3G	-0.1480	0.0105	<.0001	2.3798	BN_3G	-0.2044	0.0208	<.0001	2.3394
BN_3T	-0.1110	0.0097	<.0001	2.0434	BN_3J	0.1862	0.0199	<.0001	2.1445
BN_3H	-0.0696	0.0103	<.0001	2.2913	BN_6N	-0.1454	0.0195	<.0001	2.0674
BN_1D	0.0615	0.0097	<.0001	2.0589	BN_6F	-0.1417	0.0200	<.0001	2.1568
BN_1G	-0.0581	0.0096	<.0001	1.9851	BN_3T	-0.1268	0.0194	<.0001	2.0402
BN_6H	-0.0579	0.0097	<.0001	2.0522	BN_5I	-0.1269	0.0196	<.0001	2.0713
BN_1N	0.0529	0.0097	<.0001	2.0262	BN_5A	-0.1184	0.0193	<.0001	2.0216
BN_3E	-0.0528	0.0102	<.0001	2.2761	BN_4G	0.1151	0.0192	<.0001	2.0040
BN_3D	-0.0448	0.0098	<.0001	2.0999	BN_1D	0.1022	0.0194	<.0001	2.0378
BN_1O	-0.0431	0.0102	<.0001	2.2384	BN_1G	-0.0989	0.0191	<.0001	1.9638
BN_6L	0.0405	0.0096	<.0001	1.9842	BN_3A	0.0972	0.0195	<.0001	2.0656
BN_5J	0.0394	0.0096	<.0001	2.0064	x_{27}	0.0814	0.0190	<.0001	11.3041
BN_6F	-0.0382	0.0103	0.0002	2.2931	BN_1K	-0.0756	0.0191	<.0001	1.9781
BN_6N	-0.0371	0.0101	0.0003	2.2167	BN_1O	0.0640	0.0197	0.0012	2.0926
x_{28}	0.0312	0.0088	0.0005	10.7618	BN_4C	0.0637	0.0195	0.0012	2.0612
x_{27}	-0.0293	0.0099	0.0031	12.2065	BN_1N	0.0600	0.0194	0.002	2.0255
BN_6J	0.0278	0.0096	0.0038	1.9854	BN_3E	-0.0595	0.0205	0.0038	2.2755
BN_5A	-0.0277	0.0099	0.0053	2.1262	BN_6J	0.0554	0.0191	0.0039	1.9758
BN_4J	-0.0276	0.0100	0.0058	2.1568	BN_6L	0.0506	0.0191	0.0085	1.9829
BN_5K	0.0276	0.0104	0.0082	2.3505	x_{10}	0.0257	0.0109	0.0183	2.4092
BN_4K	-0.0227	0.0097	0.0199	2.0577	BN_6I	-0.0439	0.0207	0.0342	2.3110
BN_6I	0.0233	0.0105	0.0263	2.3685	BN_4J	0.0412	0.0197	0.0368	2.0965
BN_5H	0.0211	0.0095	0.0266	1.9608	BN_1B	0.0408	0.0198	0.0396	2.1182
x_4	0.0154	0.0070	0.0277	4.9667	BN_5H	0.0391	0.0190	0.0402	1.9567
BN_1A	0.0207	0.0097	0.0328	2.0288					
BN_3N	0.0206	0.0097	0.0332	2.0261					
BN_3A	-0.0210	0.0102	0.0397	2.2435					

on quality achieved per requiter, which does confirm some previous literature as well as intuition. The VIF for total mission per recruiter is noted; it reflects a collinearity with a non-significant term not shown in the table. This is still problematic because although the table is truncated, non-significant terms are still included in the current RMI formulation.

This issue notwithstanding, our assessment is that the RMI model with x_{29} removed is a more accurate representation of the true system—albeit with a reduced level of "fit"—and we proceed in this direction. Now that we have some idea of the current model's capabilities and limitations, we turn our attention to alternate specifications. The remaining sections of this chapter describe our results and analysis to this end.

4.3 Response Selection and First Stepwise Iteration

In the previous section, the constraints of estimating a "baseline" model dictated the use of specific independent and dependent variables. In subsequent portions of our analysis we had no such constraints and this necessitated our independent selection of appropriate response(s). This problem also relates to independent variable selection though to a much lesser extent; we have already discussed our data gathering framework and the stepwise regression procedure itself is a vehicle for suitable independent variable selection. However, neither of these is of any use if the *response* is not a suitable metric for the object of interest nor readily interpretable. In light of these considerations, we found the use of GSA_PR—from the RMI—to be somewhat counter-productive. It is not a direct measure of recruiting contracts because it is scaled by recruiter strength. Also, it is limited in scope in that it only addresses quality contracts and groups two demographics—high school seniors and older youth—together when in fact the two could respond differently to different sets of factors. Previous work which we addressed in Chapter II suggested this possibility.

We were aided in our response selection decision by the application of PCA. Since candidate responses will never be used as independent and dependent variables together in the same model, they are suitable for the application of PCA. From Table 2 and the additional definitions need for the RMI, we defined a set of seven candidate response variables, $y_1, y_2, \ldots, y_7$. Our rationale was fairly straight-forward: we included both ratios from the RMI (y_1, y_2), the raw numbers of each contract type achieved (y_3, y_4, y_5), the quality contracts achieved (y_6)—which is just y_1 less its denominator—and lastly, contract share (y_7). The results of our PCA on the candidate response set is given in Table 9.[1]

[1]The results of all our PCA are obtained at the brigade—not battalion—level with $N = 300$ (5 brigades $\times$ 60 observations). A smaller sample size allowed us to reduce the possibility of dependence between observations by aggregating the data at a higher echelon. We utilized the full set of time observations to expose as much of the data as possible as this forms the major basis of our

Table 9. PCA Summary for the Initial Response Set

	$PC_{(1)}$	$PC_{(2)}$	$PC_{(3)}$	$PC_{(4)}$	$PC_{(5)}$	$PC_{(6)}$	$PC_{(7)}$
Eigenvalues	5.2891	0.9295	0.4940	0.1626	0.1247	0.0002	0.0000
% Variance	0.7556	0.1328	0.0706	0.0232	0.0178	0.0000	0.0000
Cum. % Variance	0.7556	0.8884	0.9589	0.9822	1.0000	1.0000	1.0000
Loadings							
y_1 (GSA_PR)	**0.9336**	0.0958	-0.2768	0.1493	-0.1422	0.0092	-0.0148
y_2 (Vol_PR)	**0.9752**	0.0364	-0.0370	-0.0341	-0.2122	-0.0101	-0.0428
y_3 (GA Achieved)	**0.8970**	-0.3446	-0.2327	-0.0060	0.1498	-0.0019	-0.1105
y_4 (SA Achieved)	0.4788	**0.8702**	0.0754	-0.0001	0.0885	-0.0008	0.0020
y_5 (OTH Achieved)	**0.9156**	-0.0885	0.2510	-0.3000	-0.0271	0.0058	-0.0828
y_6 (GA+SA Achieved)	**0.9695**	0.0003	-0.1810	-0.0054	0.1649	-0.0020	-0.0979
y_7 (Contract Share)	**0.8088**	-0.1876	0.5103	0.2215	0.0345	0.0000	0.0472

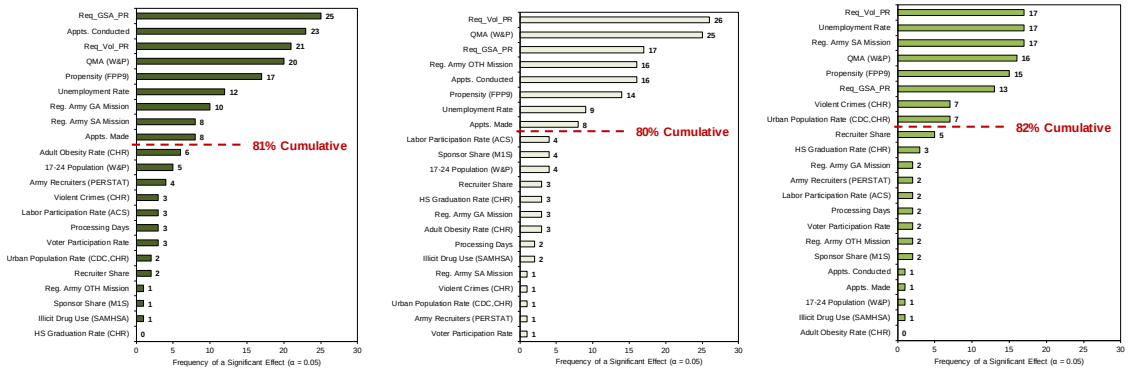
The content of Table 9 is quite interesting. The bold-face type indicates the $PC_{(i)}$ on which y_j exerts its maximum loading. The presence of bold-face type in only the first two columns of the loadings matrix indicates that seven variables are really measuring only two independent quantities. Additionally, the cumulative variance in all responses is nearly 90% by the second principal component. Practically speaking, the variables with maximum loadings on the first principal component, $PC_{(1)}$ are all highly correlated with each other but not with the sole variable making up the majority of $PC_{(2)}$. Immediately, we notice the odd variable out is that of SA contracts; this is a clear indicator that SAs should be modeled as a separate response.

Now we are faced with the choice of a response from $PC_{(1)}$. Since all six variables are approximately equally loaded, the choice might seem arbitrary. However, we have already presented our concerns regarding the use of y_1 and y_2. The same concerns apply to y_7 since it is also a ratio and is not contract type-specific. Using y_6 would be at once redundant to SA and exclusive of OTH. So, we select both y_3 and y_5 and end up with three uniformly interpretable, mutually exclusive, and collectively exhaustive metrics of AC recruiting, given our data.

Having defined appropriate responses, we then moved to an initial iteration of mixed stepwise regression. By an iteration, we mean one complete round of "stepping"

multicollinearity reduction efforts.

variables in and out of each model until none has a sufficiently (in)significant t-ratio to either enter or exit, as discussed in Chapter III. We sought to overcome a key limitation of stepwise regression—that it is not guaranteed to find a *best* subset model—by employing it separately to each battalion and then comparing the overall frequency of selected variables. This resulted in 38 different models from which to obtain our frequencies for each of the three responses. We provide the frequencies of selected variables in Figure 9(a)-(c).



(a). GA **(b). OTH** **(c). SA**

Figure 9. Frequencies of Selected Significant Variables Following First Stepwise Iteration

Sorting the frequencies in decreasing order for each contract type in Figure 9 gives them the "tornado-like" appearance. We have indicated where each of the cumulative frequencies exceeds 80% of the total. We use this 80% as a Pareto analysis rule-of-thumb (we also show in the next section a situation where Horn's criteria for PCA is met at around 80% cumulative variance). The magnitudes for GA and OTH are quite similar, as is the overall appearance and make-up of their top $\sim 80\%$. The measures of central tendency on model fits appear to indicate symmetry for all contract types but are markedly lower for SA than for GA or OTH. At this point we also note that the magnitudes for the variables in the SA models to be only about two-thirds of GA or OTH. This suggests an overall difficulty is uniquely present for fitting a SA model

to the given data.

Subsequent analysis of individual models and coefficients revealed substantial multicollinearity for all contract types. For this reason, we refrain from giving individual model parameter estimates or other diagnostics at this stage. Figure 10 demonstrates the ubiquity of the multicollinearity problem among significant terms. It is difficult to tell, either from Figure 10 or from analysis of the models themselves, exactly which terms are collinear with each other. However, the highest densities of large VIFs do appear to occur in the mission variables and mission-to-recruiter ratios. It is not surprising that these terms are collinear with each other since they are closely related. Nonetheless, the multicollinearity problem must be resolved in all models if the parameter estimates and standard errors are to be precise. In the next section, we discuss our further investigation into and resolution of multicollinearity among the independent variables.

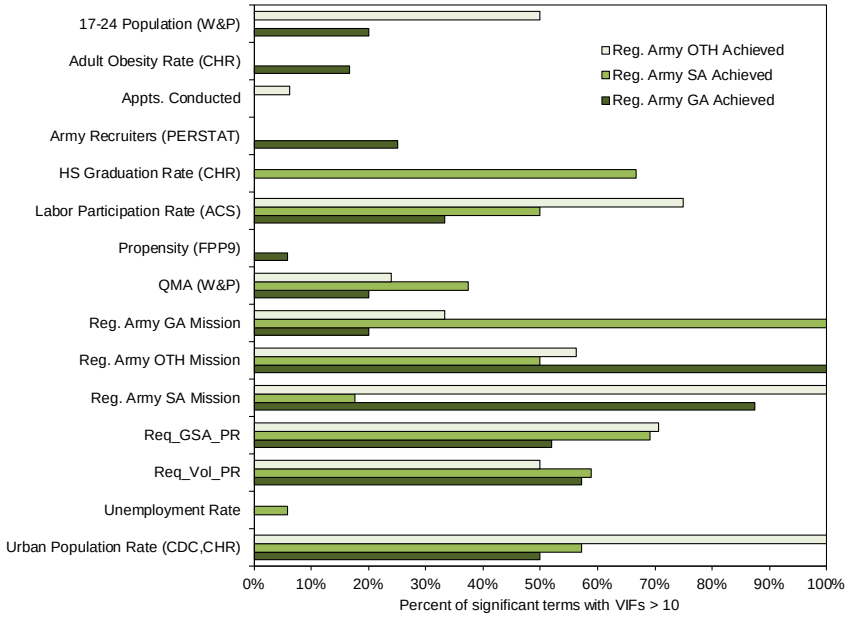


Figure 10. Variance Inflation Factors (VIF) After First Stepwise Iteration

4.4 Redefinition of Regressors and Second-Order Excursion

Following the first iteration of stepwise regression, we discovered prominent multicollinearity among the regressors of all contract types. Our feasible options for correcting this issue amounted to re-defining variables, eliminating some variables, or a combination of both.[2]. Before deciding exactly how to proceed, we applied PCA to the set of independent variables to gain insights regarding their variance structure. Figure 11 provides the sorted eigenvalues of each component and Horn's curve; Table 10 provides the data and the loadings matrix for all retained components (i.e., all components with eigenvalues greater than Horn's curve).

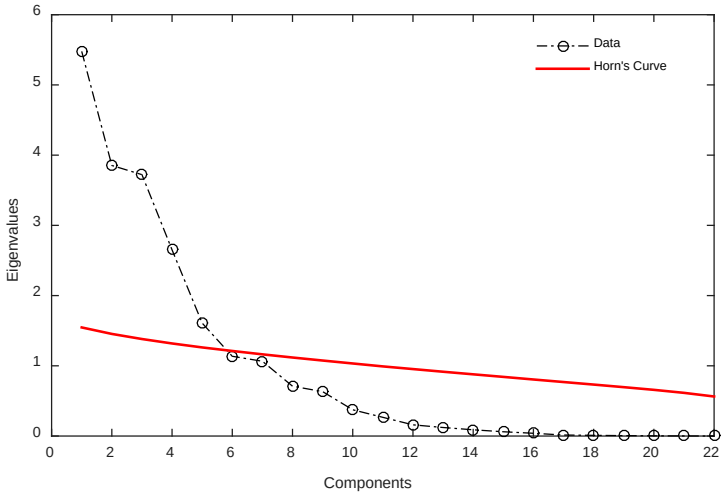


Figure 11. PC Eigenvalues and Horn's Curve for the First Set of Regressors

From Figure 11 and Table 10, the five retained components account for about 79% of the variance in the original 22 variables. The GA and OTH missions, along with propensity among others, are loaded most heavily on the first component. The second component consists mostly of several lossely connected demographic factors. The third component is of particular interest as it contains both the mission ratio terms but also the recruiter strength and SA mission. The fourth and fifth components

[2]Collecting additional data and other complex methods such as principal components regression are also valid approaches to combating multicollinearity [43], but these were beyond our scope.

Table 10. PCA Summary for the Initial Independent Variable Set

	$PC_{(1)}$	$PC_{(2)}$	$PC_{(3)}$	$PC_{(4)}$	$PC_{(5)}$
Eigenvalue	5.4776	3.8533	3.7221	2.6511	1.6084
% Variance	0.2490	0.1752	0.1692	0.1205	0.0731
Cum. % Variance	0.2490	0.4241	0.5933	0.7138	0.7869
Loadings					
x_1 (Voter Part. Rate)	0.4724	**0.6486**	-0.0628	-0.0948	0.2746
x_2 (Sponsor Share)	0.5045	**0.7153**	0.0180	0.0579	-0.0407
x_3 (Labor Part. Rate)	**-0.7619**	0.2076	0.2354	-0.3092	0.2346
x_4 (Unempl. Rate)	-0.0377	-0.4225	**-0.6136**	-0.0599	0.5270
x_5 (HS Grad. Rate)	-0.4288	**0.5901**	0.4573	0.0041	-0.1782
x_6 (Violent Crimes)	-0.0709	-0.2936	**-0.7065**	-0.3887	0.0780
x_7 (Obesity Rate)	0.5014	**0.7327**	0.0519	0.1881	-0.0123
x_8 (Drug Use Rate)	**-0.7177**	-0.4122	0.0862	0.1361	0.2085
x_9 (Urban Pop. Rate)	-0.4932	**-0.7896**	0.0706	-0.0925	-0.2529
x_{10} (Propensity)	**0.6309**	-0.4604	-0.2794	0.2299	-0.3635
x_{11} (QMA Pop.)	**-0.6595**	0.3620	0.2762	-0.4695	0.1842
x_{12} (17–24 Pop.)	**-0.7745**	0.1379	0.3073	-0.4403	0.0694
x_{15} (GA Mission)	**0.5548**	-0.2439	0.2772	-0.4699	0.4496
x_{16} (SA Mission)	0.2977	-0.1260	**0.7181**	0.1922	0.1013
x_{17} (OTH Mission)	**0.7237**	-0.2960	-0.0005	-0.2013	0.2052
x_{22} (Recruiter Share)	0.3273	0.1584	-0.2562	**-0.6416**	-0.0017
x_{23} (PERSTAT Recruiters)	0.3109	0.2602	**-0.5920**	-0.5654	0.0165
x_{24} (Appts. Made)	0.2679	-0.1399	0.3664	**-0.7385**	-0.3609
x_{25} (Appts. Cond.)	0.3141	-0.2481	0.3994	-0.5156	**-0.5592**
x_{26} (Process. Days)	-0.0334	0.0158	0.0654	**0.1292**	-0.0716
x_{27} (Req_Vol_PR)	0.5856	-0.3851	**0.6286**	-0.0026	0.3038
x_{28} (Req_GSA_PR)	0.3848	-0.2956	**0.7824**	0.0378	0.2950

have relatively light loadings but it is interesting to note that appointments made and appointments conducted are not under the same component as might be expected. Only the exact linear combinations for each principal component are orthogonal, but a judicious choice of variables from among the five retained components should result in a much reduced set of regressors that are minimally correlated. A prudent selection of variables will consider the top portions of the tornado charts from Figure 9, as well as diversification from among the principal components. In other words, we wish to retain as much original information as possible while minimizing collinearity.

This suggests a strategy of somehow combining variables which load on the same principal component, if indeed their inclusion is warranted and their redefinition is interpretable. Common threads from the tornado charts are the inclusion of unemployment rate, propensity, appointments made and conducted, QMA, as well as both missions and mission-to-recruiter ratios. However, just looking at the first component

reveals loadings of four of these variables. Of these, we place primary importance on the missions for GA and OTH due to the reliability of the data, their direct control by USAREC, and previous findings regarding their relative importance. Therefore, we should look for another way to express propensity and QMA so that they are not correlated with the missions. Incidentally, none of the variables loading on $\text{PC}_{(2)}$ figure prominently in the tornado charts although from Chapters I and II, both obesity and high school graduation rates are thought to play a role in the qualification status of potential Soldiers. If we can redefine a variable that captures information both on propensity and qualification of youth to serve, we may be able to place this new variable in the "vacant" space of the orthogonal second principal component and then retain that variable for further use.

In light of this idea, we propose a new variable (x_{33}) defined as the number of 17–24 year old youths who are jointly probable—assuming independence—to be in the potential, target, and QMA markets (PTQMA). We use the definitions of these respective markets given by USAREC 3-0 to establish the following events:

- $P =$ someone is in the potential market (i.e., propensed); $P \equiv x_{10}$
- $T =$ someone is in the target market (i.e., high school diploma graduate); $T \equiv x_5$
- $Q =$ someone is qualified (i.e., is physically fit); $Q \equiv 1 - x_7$

In reality, our definition of these events amounts to some very broad assumptions; these metrics are likely far too limited to be considered accurate in any real sense. For example, the target market doctrinally considers only male graduates with high aptitudes (we have only represented graduation rate without respect to gender). Obviously, qualification amounts to much more than lack of obesity, although this is all we can represent. However, our primary purpose is achieved in that we can now retain information which is independent and has proven to have at least some importance in a predictive model. Using x_{12} for the youth population and the events just described,

we complete our definition of PTQMA as

$$\begin{aligned} x_{33} &= x_{12}PTQ \\ &= x_{12}(x_{10})(x_5)(1 - x_7) \end{aligned} \tag{35}$$

which is completed for all units and time periods (scripts omitted). We add PTQMA to the original regressors and re-compute the principal components as given in Figure 12 and Table 11.

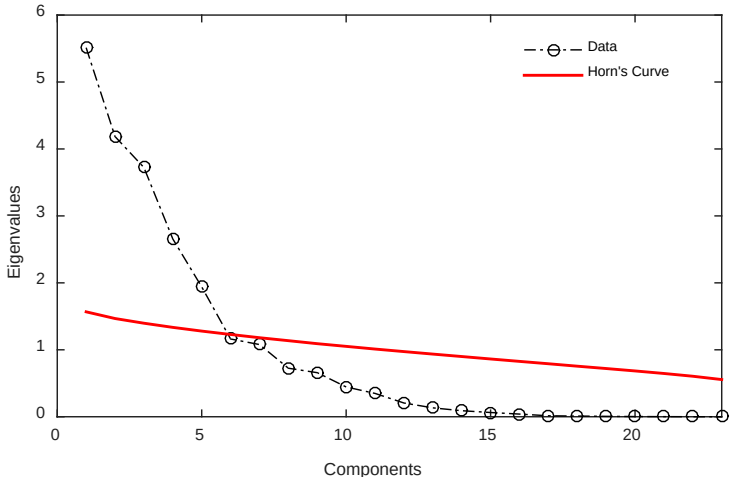


Figure 12. PC Eigenvalues and Horn's Curve for the First Set of Regressors + PTQMA (x_{33})

From the second PCA, the new variable x_{33} is clearly loaded on the second principal component, while GA and OTH missions remain on the first. Appointments made and conducted are now (more intuitively) aligned together on the fourth component, with unemployment the only significant loading on the fifth component. Therefore, we propose several further redefinitions. First, we create a ratio of appointments conducted to appointments made (x_{30}, C-M Ratio). This combines two elements of the fourth component and does it in such a way so as to conform to the logical recruiting sequence. Second, add the GA and OTH missions together to form a single mission (x_{31}, GA+OTH Msn). These are common elements of the first component but there

Table 11. PCA Summary for the Initial Independent Variable Set + PTQMA (x_{33})

	$PC_{(1)}$	$PC_{(2)}$	$PC_{(3)}$	$PC_{(4)}$	$PC_{(5)}$
Eigenvalue	5.5128	4.1851	3.7227	2.6558	1.9398
% Variance	0.2397	0.1820	0.1619	0.1155	0.0843
Cum. % Variance	0.2397	0.4216	0.5835	0.6990	0.7833
Loadings					
x_1 (Voter Part. Rate)	0.4129	**0.7236**	-0.0974	-0.0905	-0.2317
x_2 (Sponsor Share)	0.4588	**0.7160**	-0.0227	0.0857	0.1995
x_3 (Labor Part. Rate)	**-0.7789**	0.1775	0.2229	-0.3025	-0.0876
x_4 (Unempl. Rate)	-0.0234	-0.3727	-0.5865	-0.0983	**-0.5899**
x_5 (HS Grad. Rate)	-0.4578	**0.5183**	0.4211	0.0408	0.3632
x_6 (Violent Crimes)	-0.0497	-0.3171	**-0.6883**	-0.4013	-0.0862
x_7 (Obesity Rate)	0.4457	**0.7690**	0.0111	0.2069	0.0515
x_8 (Drug Use Rate)	**-0.6945**	-0.4229	0.1094	0.1153	-0.2744
x_9 (Urban Pop. Rate)	-0.4277	**-0.8455**	0.1137	-0.1051	0.1460
x_{10} (Propensity)	**0.6840**	-0.5078	-0.2544	0.2324	0.3201
x_{11} (QMA Pop.)	**-0.6923**	0.3585	0.2555	-0.4614	-0.0954
x_{12} (17–24 Pop.)	**-0.7871**	0.1130	0.2986	-0.4340	-0.0119
x_{15} (GA Mission)	**0.5567**	-0.1272	0.2929	-0.4931	-0.4036
x_{16} (SA Mission)	0.3015	-0.0598	**0.7245**	0.1859	-0.1073
x_{17} (OTH Mission)	**0.7331**	-0.1972	0.0183	-0.2235	-0.2657
x_{22} (Recruiter Share)	0.3152	0.1751	-0.2641	**-0.6357**	0.0660
x_{23} (PERSTAT Recruiters)	0.2959	0.2457	**-0.6053**	-0.5558	0.0807
x_{24} (Appts. Made)	0.2792	-0.1128	0.3737	**-0.7304**	0.3220
x_{25} (Appts. Cond.)	0.3383	-0.2395	0.4121	**-0.5050**	0.4637
x_{26} (Process. Days)	-0.0322	0.0052	0.0640	**0.1325**	0.0629
x_{27} (Req_Vol_PR)	0.5980	-0.2586	**0.6513**	-0.0282	-0.3534
x_{28} (Req_GSA_PR)	0.3932	-0.1871	**0.7993**	0.0184	-0.3057
x_{33} (PTQMA)	0.2259	**-0.6351**	-0.0145	0.0783	0.5861

is no immediately relevant reason for a ratio. Third, create a ratio of the SA mission to recruiters (x_{32}, Req_SA_PR). This last ratio consists of common, relevant elements of the third component, and the only logical way to combine them is with a ratio. We provide the explicit definitions of x_{30} through x_{32} in equation (36).

$$\begin{aligned} x_{30} &\equiv \frac{z_{25}}{z_{24}} = \text{appointments conducted to made (C-M ratio)} \\ x_{31} &\equiv z_{15} + z_{17} = \text{GA+SA mission (GA+OTH Msn)} \\ x_{32} &\equiv \frac{z_{16}}{z_{23}} = \text{total contracts achieved per recruiter (Req\_SA\_PR)} \end{aligned} \tag{36}$$

With this set of variables, we capture the relevant pieces of each of the five principal components while minimizing the duplication of information. We discard the remaining variables either because they were not relevant to the models following the

first stepwise iteration, or because they are already captured as part of one of our five retained variables. In Table 12, we provide the correlation matrix, **R**, for the five retained regressors. Of the ten off-diagonal ($r_{i,j}$) elements, seven are less than |0.2|. We do note that $|r_{4,32}| = 0.37$ and $|r_{30,33}| = 0.41$ but these are still both less than 0.5 and are therefore not overly troublesome. Therefore, we are confident that the reduced set of five variables—which actually retains information from 12 original variables—will be adequate to reduce multicollinearity going forward.

Table 12. Correlation Matrix R for the Reduced Set of Independent Variables

	x_4	x_{30}	x_{31}	x_{32}	x_{33}
x_4 (Unempl. Rate)	1	-0.2133	0.1964	-0.3698	-0.0863
x_{30} (C-M Ratio)		1	-0.1299	0.1995	0.4095
x_{31} (GA+OTH Msn)			1	0.0732	0.0788
x_{32} (Req_SA_PR)		*symm.*		1	0.0772
x_{33} (PTQMA)					1

As a concluding step to this portion of our research, we completed a second iteration of mixed stepwise regression for the individual unit models. However, this time we used the 5-regressor set, replete with second-order terms and all possible two-way interactions. As discussed in Chapter III, this constitutes our attempt to test for non-linearities and interactions as part of a second-order response surface model. We had some reason to believe that either x_{31} or x_{32} might be non-linear—specifically, concave down—since these terms can be thought of as measuring recruiter effort as pointed out in Chapter II. However, our results do not appear to indicate recurring significance of any non-linear effects as shown in Figure 13.

We have now completed the initial stages of the model-building process. These were also the most complex, as they required us to build from the ground up. However, we now have well-defined responses and a parsimonious set of independent variables that should have adequate predictive power in addition to low collinearity. In the next section, we discuss our final model specification and adequacy-checking procedures.

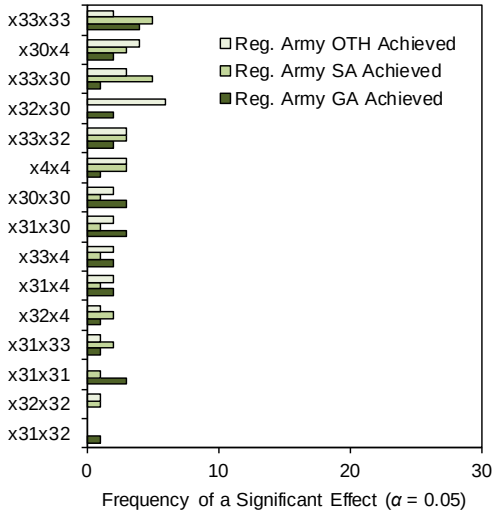


Figure 13. Frequencies of Significant 2nd Order Response Surface Terms Following Second Stepwise Iteration

4.5 Final Model Specification and Adequacy

From the discussion in Chapter III, a single model with several indicator variables is equivalent to a different model for each indicator. We make use of this feature here during our final iteration of stepwise regression for a few reasons. First, it enhances adequacy-checking by dramatically increasing the sample size. Second, it will be easier to make an assessment of the model's overall fit since this will be codified in a single ANOVA and set of summary statistics. Finally, this type of specification is likely to be more parsimonious since the differences between battalions may be more prevalent than differences within each unit over time. If this is the case—as we suspect based on the relatively low R^2 values thus far—then this can be modeled by allowing intercepts (and slopes) to change between units in a single model. Therefore instead of fitting a unique model to every single battalion, we allowed for the interaction of all battalions with each main effect term during the stepwise selection procedure. This is equivalent to allowing the slopes to change for every battalion. We obtain the summary statistics for each model in Table 13. Clearly the regressions are significant

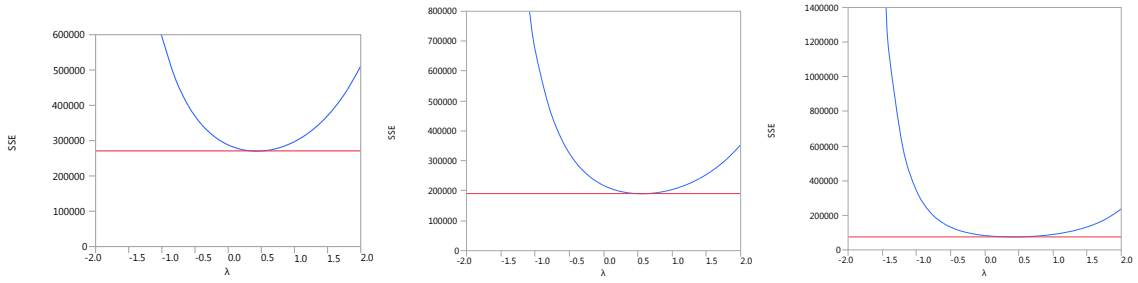
for each contract type.

Table 13. Summary of Fit for the Final Models Selected Following Third Stepwise Iteration

Response, $y^{(k)}$	R^2_{Adj}	R^2_{Pred}	$P > F_0$	p	N
Reg. Army GA Achieved, $y^{(GA)}$	0.7368	0.7261	< 0.0001	62	1710
Reg. Army SA Achieved, $y^{(SA)}$	0.5451	0.5254	< 0.0001	50	1710
Reg. Army OTH Achieved, $y^{(OTH)}$	0.7926	0.7832	< 0.0001	63	1710

Adequacy.

Given that the RMI baseline models showed some signs of heteroskedasticity, we first examined the recommended Box-Cox transformations on each $y^{(k)}$ in Figure 14. For all three contract types, the SSE is minimized in the vicinity of $\lambda = 0.5$, which indicates a square root transformation.



(a). GA **(b). OTH** **(c). SA**

Figure 14. Box-Cox Transformations for $y^{\lambda(k)}$

After applying the transformation $y' = \sqrt{y}$ to correct for non-constant variance, we then examined the independence of the data using the Durbin-Watson (DW) test. However, we were not able to use the DW test statistic from JMP® since our data was divided by categorical variables. Therefore, we performed the DW test manually on each sequence of observations within every battalion, using original code in MATLAB. Figure 15 displays values of d and $4-d$, the DW test statistics for positive and negative autocorrelation, respectively. On the left-hand side, we show the test statistics for

the current transformed model. It is clear that we have several units with some autocorrelation or inconclusive results in every contract type (using the bounds for $N = 40$ or $N = 50$ is somewhat subjective). On the right side of Figure 15, we modify the transformed model to include $\phi_i^{(k)}$, a lag-1 autocorrelation parameter for each i battalion and each k contract type. By including $\phi_i^{(k)}$, we have now brought every unit out of the rejection region for $N = 40$. This effect is particularly pronounced for SA contracts. For the scope of this project, we assume an inconclusive result to be satisfactory. Thus, we will continue to leave the autocorrelation parameters in the model for all units irrespective of significance; this is an unfortunate added complexity but is necessary to maintain adequacy assumptions.

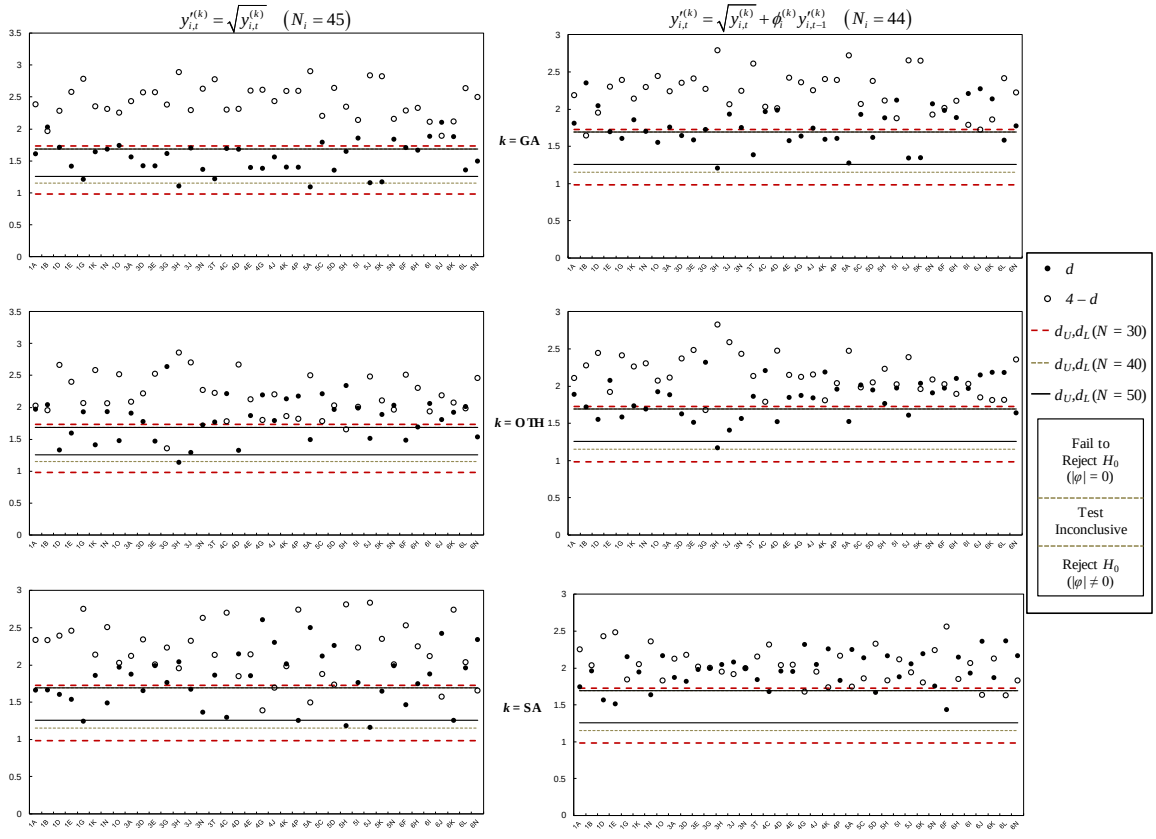


Figure 15. Result of Tests for Autocorrelation for $y_i'^{(k)}$ without (left) and with (right) $\phi_i^{(k)}$

In the case of SA contracts, we also observed evidence that time-dependency was highly correlated with period of the year. This can easily be seen in the time series plot of SA production in Appendix D. Given the relatively poor performance of the SA model thus far, we conducted an excursion by introducing an additional series of indicators for quarters of the recruiting year. We used quarter 1 as the baseline (intercept) and allowed these terms to interact with each unit during an additional iteration of stepwise regression.

Our results indicated all three quarters to have significant effects and the fit of the SA model was improved dramatically. Table 14 contains the final summaries of fit for the transformed models with lag-1s and quarter indicator variables for SA contracts. Comparisons between the values of p, the number of explanatory variables plus the intercept, in Tables 13 and 14 should be made with care. While for each contract type k we have added an additional 38 parameters for autocorrelation (and quarters for SA), we have also removed several non-significant terms which are also not required by the hereditary rule.

Table 14. Summary of Fit for the Final Transformed, Lag-1 Models with Non-significant, Non-hereditary Terms Removed

Response, $y^{(k)}$	R^2_{Adj}	R^2_{Pred}	$P > F_0$	p	N
(Reg. Army GA Achieved)$^{1/2} = y'^{(GA)}$	0.7402	0.7289	< 0.0001	89	1672
(Reg. Army SA Achieved)$^{1/2} = y'^{(SA)}$	0.6983	0.6794	< 0.0001	100	1672
(Reg. Army OTH Achieved)$^{1/2} = y'^{(OTH)}$	0.8069	0.7954	< 0.0001	98	1672

In Table 15, we provide a summary of potential leverage and influential points. The table reflects the special attention we paid to points with large values of h_{ii} or r_i. We truncate time column by the convention, $mmyy$. While the criteria for a large hat diagonal is straightforward ($> 2p/n$), what constitutes a large residual is more subjective. We chose to apply a conservative rule of 5%, or the upper- and lower- 2.5% of the distribution of r_i. We show in bold text intersections of large hat

diagonals and large internally studentized residuals. There is only one leverage point (unusual in the input-space) for each contract type, but the lack of any large values of D_i indicates that there do not exist any influential points.

Table 15. Leverage and Influence Data for the Transformed, Lag-1 Models

$y'^{(GA)} = \sqrt{y^{(GA)}} + \phi y'$					$y'^{(SA)} = \sqrt{y^{(SA)}} + \phi y'$					$y'^{(OTH)} = \sqrt{y^{(OTH)}} + \phi y'$				
Unit	Time	r_i	h_{ii}	D_i	Unit	Time	r_i	h_{ii}	D_i	Unit	Time	r_i	h_{ii}	D_i
1A	1006	-1.68	0.11	0.00	1N	1209	2.19	0.13	0.01	1K	1303	-1.67	0.14	0.00
1K	1309	-1.86	0.14	0.01	1N	1210	1.92	0.13	0.01	1K	1305	2.26	0.22	0.01
1O	1012	2.68	0.15	0.01	1O	1106	2.33	0.18	0.01	3N	1009	1.67	0.12	0.00
3A	1010	2.66	0.12	0.01	1O	1204	-1.89	0.13	0.01	3N	1301	2.02	0.14	0.01
3A	1011	1.69	0.15	0.01	1O	1309	-2.54	0.13	0.01	3T	1204	-2.58	0.20	0.02
3A	1012	2.64	0.16	0.01	3J	1307	-1.67	0.15	0.00	5C	1111	**3.50**	**0.33**	0.06
3A	1101	**-2.12**	**0.24**	0.02	3N	1111	1.95	0.19	0.01	5C	1204	-1.91	0.18	0.01
3A	1105	1.71	0.11	0.00	3N	1208	2.17	0.16	0.01	5I	1305	-1.73	0.14	0.00
3A	1204	-2.08	0.15	0.01	3N	1309	-2.56	0.15	0.01	5I	1307	2.26	0.13	0.01
3D	1105	1.97	0.11	0.01	4C	1208	2.51	0.13	0.01	6K	1101	-1.95	0.12	0.01
5A	1209	-1.90	0.13	0.01	4E	1010	1.81	0.16	0.01	6K	1307	2.12	0.14	0.01
5I	1006	-1.91	0.11	0.00	4J	1209	**-2.09**	**0.33**	0.02	6L	1007	-1.76	0.16	0.01
6I	1011	1.74	0.12	0.00	4P	1208	2.25	0.17	0.01	6N	1104	-1.86	0.14	0.01
6I	1012	2.69	0.12	0.01	4P	1305	1.68	0.15	0.00	6N	1308	1.83	0.19	0.01
6J	1008	-2.08	0.16	0.01	4P	1308	1.79	0.13	0.00					
6N	1012	1.89	0.11	0.00	4P	1309	-3.86	0.15	0.03					
					5H	1308	3.19	0.17	0.02					
					5J	1003	-1.87	0.17	0.01					
					5J	1010	1.79	0.13	0.00					
					5J	1309	-2.29	0.14	0.01					
					6I	1007	-1.65	0.16	0.00					
					6I	1109	-3.16	0.13	0.01					
					6K	1208	1.64	0.18	0.01					

In fact, for all contract types D_{max} remained less that 0.25, well below the recommended criteria of unity. The bold points are also labeled in the subsequent plots for identification. We do notice the prominence of certain units such as 3A and 4P, as well as the trend that potential leverage points for a given unit generally lie close together in time. This is likely a function of an un-modeled time dependency, although we do not venture any other insight. Since the bold points are few and not influential, we also do not see a reason to exclude them from the model. However, it may be prudent in the future for USAREC or other research to verify the conditions surrounding these data points.

At this point, adequacy checking is complete. In Figure 16(a)-(f), we present the final normal quantile and predicted plots for the internally studentized residuals. It is

apparent that the adequacy assumptions for normality and constant variance are satisfied, following transformation of the responses and the inclusion of a lag-1 response. Pairs (a,b), (c,d) and (e,f) show the plots for GA, SA, and OTH, respectively. Note that with the brigade color scheme applied to each plot, for OTH contracts there is fairly clear delineation between the predicted values for the 2nd Brigade and the 1st/3rd Brigades. The predicted OTH contracts for 5th and 6th Brigades appear to be evenly distributed. Therefore, Figure 16f would appear to suggest that a majority of the predicted non high-quality contracts originate in the southeast region of 2nd Brigade. This distinction appears to be less prominent for GA and SAs, suggesting those contract types are predicted to be more evenly dispersed between regions.

Model Forms.

Now that we have completed adequacy checking, we are able to make inferences regarding the coefficients. Therefore, we conclude this section with a presentation of the adequate model coefficients and offer several interpretations of the final model forms. We begin with Tables 16 and 17, which contain the coded coefficients for the main model parameters of each contract type. We begin with the coded coefficients since they are useful for comparisons of relative importance between the terms. First, we note that all VIFs are well below 10 as is recommended [43]. Also, we see that all main effect terms are significant with one exception each for OTH and SA (these two terms remain in the models due to significant interactions with unit indicators, which are not shown here but addressed in following discussion).

For GA and OTH, the total mission of these two combined categories has about three times the positive impact on contract production than does unemployment. This is broadly consistent with previous findings. Increasing the SA mission per recruiter appears to decrease GA and OTH contracts, although this effect is only

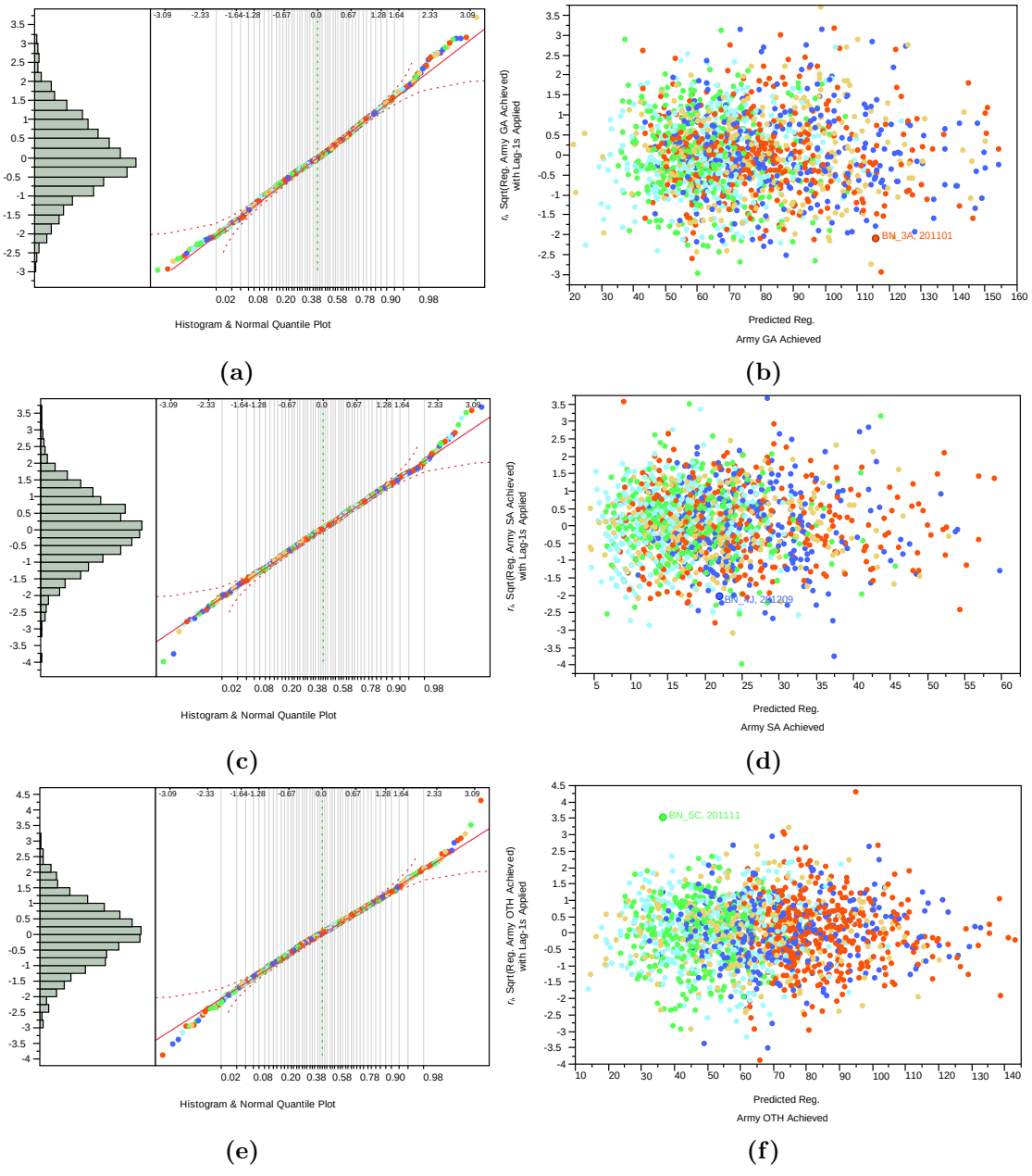


Figure 16. Final Quantile Plots and Residual (r_i) Plots of the Adequate Models

Table 16. Main Effect Coefficients in Coded Units for $y_t'^{(k)} = \sqrt{y_t^{(k)}} + \phi^{(k)} y_{t-1}'^{(k)}$

	$k = GA$					$k = OTH$				
x_j	$b_j^{(k)}$	se$(\cdot)$	t_0	$P > \|t_0\|$	VIF	$b_j^{(k)}$	se$(\cdot)$	t_0	$P > \|t_0\|$	VIF
Intercept	9.1169	0.1092	83.49	0.0001		8.3089	0.1020	81.48	0.0001	
x_4 (Unemp. Rate)	1.2023	0.1310	9.18	0.0001	4.11	0.7719	0.1403	5.50	0.0001	6.19
x_{30} (C-M Ratio)	0.2043	0.1012	2.02	0.0438	1.98	−0.2830	0.0941	−3.01	0.0027	2.28
x_{31} (GA+OTH Msn)	3.0943	0.1232	25.12	0.0001	3.61	2.5464	0.1135	22.44	0.0001	3.63
x_{32} (Req_SA_PR)	−0.7111	0.1177	−6.04	0.0001	2.39	−0.0962	0.1086	−0.89	0.3761	2.41
x_{33} (PTQMA)	0.8835	0.1531	5.77	0.0001	4.77	0.3672	0.1424	2.58	0.0100	4.40

Table 17. Main Effect Coefficients in Coded Units for $y_t'^{(SA)} = \sqrt{y_t^{(SA)}} + \phi^{(SA)} y_{t-1}'^{(SA)}$

x_j	$b_j^{(SA)}$	se$(\cdot)$	t_0	$P > \|t_0\|$	VIF
Intercept	4.0833	0.1119	36.4900	0.0001	
x_4 (Unemp. Rate)	−0.5137	0.1093	−4.70	0.0001	3.74
x_{30} (C-M Ratio)	0.0934	0.0898	1.0400	0.2981	1.96
x_{31} (GA+OTH Msn)	1.3000	0.0845	15.38	0.0001	2.15
x_{33} (PTQMA)	0.0888	0.1304	0.68	0.4958	4.50
QTR2	0.3167	0.0470	6.74	0.0001	1.67
QTR3	0.8169	0.0482	16.96	0.0001	1.59
QTR4	−0.3661	0.0499	−7.34	0.0001	1.62

statistically significant as a main effect for GA. This suggests a palpable trade-off for recruiters between pursuing GA and SA contracts. However, the largest enhancer of SA contracts is also apparently the GA+OTH mission which suggests that recruiters may be more likely to convert a SA mission into a GA contract than vise-versa. We wonder if this is due to recruiter perception about the lack of a SA market, or a hidden organizational incentive. Finally, we observe that the C-M Ratio has a negative effect for OTH contracts while for GA it is positive. While merely speculative, we might possibly attribute this to a competition for conducted appointments between GA and OTH. In such a scenario, an increased C-M ratio might decrease production of OTH contracts if indeed the increased ratio occurs as a result of conducting more appointments with GA prospects than with OTHs. This is plausible given USAREC's emphasis in recruiting quality contracts.

Now regarding the SA model: interestingly, the Req_SA_PR term is not significant and not involved in any interactions in the SA model. This seems counter-intuitive,

but may very well be explained by the high degree of significance shown by the quarter indicators. We must conclude based on the sample data that the SA mission-to-recruiter ratio does not affect SA contracts. In other words, the regularity of the cyclical fluctuations in achieved SA contracts may overwhelm any effect attributed to the mission per recruiter ratio. In any case, another counter-intuitive finding for the SA model is that unemployment appears to decrease the number of SA contracts (the opposite is true for GA and OTH). This is quite puzzling in light of previous findings regarding the positive correlation of unemployment with SA contract production. However, we also note the sensitivity of time to changes in behavior and acknowledge that this may be a factor. We can only speculate, but offer that perhaps the current generation of senior youth is less driven to pursue an Army career if—being more susceptible to peer pressure than their older counterparts—they are influenced by less productive behavior in their local area. A more plausible explanation can possibly be found by comparing the two time series directly, whereupon one finds that cyclical lows in SA production correspond almost exactly with seasonal highs in unadjusted unemployment. We do not speculate whether or not this relationship is coincidental, except to say that it could provide a sensible explanation for the contrary sign of unemployment with respect to this specific contract type.

As we have just shown, the coded coefficients are useful to draw interpretive conclusions. However, the natural coefficients are needed in practice to implement the model. Therefore, we provide the main effects in natural units in Table 18. The contents of Figure 18 can be obtained directly from JMP®, as can details on the individual parameters that are unit-specific. In Figure 17 we provide an example excerpt of the software output for 3 unit-specific terms out of the 89 total terms estimated in the GA model.

Unfortunately, the software does not complete the required algebraic operations

Table 18. Main Effect Coefficients in Natural Units for $y_t'^{(k)} = \sqrt{y_t^{(k)}} + \phi^{(k)} y_{t-1}'^{(k)}$

ξ_j	$\hat{\beta}_j^{(GA)}$	$\hat{\beta}_j^{(SA)}$	$\hat{\beta}_j^{(OTH)}$
Intercept	2.1978	3.2325	4.8331
ξ_4 (Unemp. Rate)	26.4225	−12.7345	15.2018
ξ_{30} (C-M Ratio)	0.8758	0.5186	−0.9503
ξ_{31} (GA+OTH Msn)	0.0245	0.0107	0.0200
ξ_{32} (Req_SA_PR)	−2.0811	n.s.	−0.2294
ξ_{33} (PTQMA)	10.58×10^{-6}	1.47×10^{-6}	4.82×10^{-6}
QTR2	n.s.	0.3442	n.s.
QTR3	n.s.	0.8373	n.s.
QTR4	n.s.	−0.3518	n.s.

Parameter	Beta (β)	Std Beta (b)	Std Error	t Ratio	Prob>|t|	VIF
BN_3N	-1.7720	-0.1024	0.1519	-0.6700	0.5005	2.6346
Unemployment Rate*BN_3N	20.9050	1.2509	0.5301	2.3600	0.0184	2.6057
GA Lag1*BN_3N	-0.0014	0.1332	0.5548	0.2400	0.8102	8.4011

Figure 17. JMP® Output Example for Three Terms Specific to Battalion 3N (Tampa)

to combine unit-specific terms and main effects to create individual unit models. As we discussed in Chapter III, the individual battalion values for each coefficient may differ, indicating a unit-specific intercept and/or slope coefficient(s). Therefore, we must manually create individual battalion models and now briefly illustrate our procedure for accomplishing this. To begin, we write out the full model using Table 18 as

$$\hat{y}_t^{(i)} = 2.20\xi_{0,t}^{(i)} + 26.42\xi_{4,t}^{(i)} + 0.88\xi_{30,t}^{(i)} + 0.02\xi_{31,t}^{(i)} - 2.08\xi_{32,t}^{(i)} + 1.1\text{E}^{-6}\xi_{33,t}^{(i)} \tag{37}$$

for $i = 1\text{B}, 1\text{D}\ldots, 6\text{N}$. Once we select $i = 3\text{N}$, we scan the software output for any terms containing this indicator and find the three terms indicated by Figure 17. Then from equation (11), these terms—provided they are significant or required for heredity or

adequacy assumptions—are added to their respective coefficients in (37):

$$\begin{aligned}\hat{y}_t^{(3N)} &= (2.20 - 1.77)\xi_{0,t}^{(3N)} + (26.42 + 20.91)\xi_{4,t}^{(3N)} + 0.88\xi_{30,t}^{(3N)} + 0.02\xi_{31,t}^{(3N)} \\ &\quad - 2.08\xi_{32,t}^{(3N)} + 1.1\text{E}^{-6}\xi_{33,t}^{(3N)} + (0.0057 - 0.0014)\phi\hat{y}_{t-1}^{(3N)} \\ &= 0.43\xi_{0,t}^{(3N)} + 47.33\xi_{4,t}^{(3N)} + 0.88\xi_{30,t}^{(3N)} + 0.02\xi_{31,t}^{(3N)} - 2.08\xi_{32,t}^{(3N)} \\ &\quad + 1.1\text{E}^{-6}\xi_{33,t}^{(3N)} + 0.0043\phi\hat{y}_{t-1}^{(3N)}\end{aligned} \tag{38}$$

In this case not all of the terms changed; this indicates statistical non-significance of the unit's slopes for x_{31}, x_{32}, and x_{33} from the baseline unit, 5N. In Tables F.1, F.3, and F.2, we give the models for all individual units which include the contents of Table 18 as well as the intercept shifts and slope changes necessitated by the significant unit-specific terms. Note that for most units, non-intercept terms tend to not differ significantly or with any pattern from the baseline. This appears to indicate that effects of these variables are largely fixed between units. Therefore, it is possible that differing slopes—in the few cases where they do occur—may be fitting more noise than signal. This concludes our final section with regards to model construction. We now move to the closing section of the chapter as we assess the performance of our models against the validation data set.

4.6 Validation

Having formulated and analyzed a set of adequate models, we now seek to test their stability—and consequently their utility—using some new data. As mentioned in Chapter III, we set aside the last 15 months of data for this exact purpose. Now because our data is cross-sectional with observations taken at each points in time across multiple categories, we have multiple options for conveying our results. However, it important to note how the sample size changes depending on our object(s) of comparison. For example, a single categorical entity (i.e., battalion or USAREC-

aggregate) over time has sample sizes $N_E = 44$ and $N_V = 15$ for estimation and validation datasets, respectively; we denote a comparison of this type by the term *echelon*. However, a single comparison made between all recruiting battalions over time has sample sizes $N_E = 1672$ and $N_V = 570$ since time periods must be multiplied by the number units; we refer to this type as a *comprehensive* comparison. When summarizing a group of echelon comparisons we provide the comprehensive average as a baseline for comparison as opposed to averaging the individual echelon values; this is an effort to reduce bias that might be introduced from "averaging the averages."

We begin our validation analysis at the macro- (HQ USAREC) level and work our way down to specific contract types and battalions. We start by presenting a single time series chart for total contract production in Figure 18. The gray line represents the actual data, while the solid black line represents our predictions for the estimation dataset. For the validation period, which runs from the tenth month of RY13 to the last month of RY14, we include predictions and one-period ahead prediction intervals (PI) for 80% and 90% confidence, respectively. The narrower PI assumes that the inputs are known exactly (i.e., with perfect clairvoyance); the wider interval assumes that USAREC's inputs are known, but that unemployment and PTQMA are not known and are therefore forecast for each battalion using simple linear trend models. Both known and unknown PIs are respectively symmetric about their predicted data, although the reader will note non-symmetry of these PIs to each other. The latter phenomenon is attributable to the forecast inputs differing from known inputs, which in turn produce different predicted data. For brevity, we provide the predicted data obtained from known inputs with a thick dashed line.

From Figure 18, our predictions explain 72% of the validation data variation at the HQ USAREC level. In general, we consistently capture the fluctuations of the

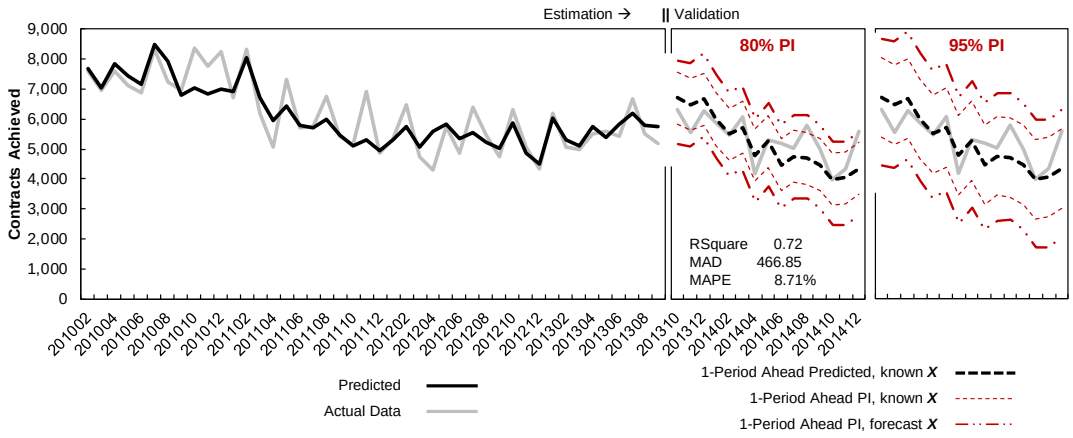


Figure 18. Contracts Achieved and Model Predictions, HQ USAREC-echelon Total Over All Contract Types

estimation data, although there appears to be some indication that upward spikes are not as precisely fit. Also, we note a three-month period between months tenth and twelfth months of RY10 where our predictions underestimate the actual data by as much as 1,346 contracts per month; this is quite a large deviation considering the estimation data MAD for total contracts is only 424 contracts per month. Therefore, it could be possible that a departure from the underlying process of contract production occurred during this three-month span.

Continuing our analysis, we note a mild tendency to underestimate total production in the latter parts of the validation period. This is to be expected to some extent, given increasing distance from the forecast origin. At any rate, the 80% PI assuming forecast inputs appears to be quite adequate in the characterization of error. Table 19 states response-unit validation PIs for both 80% and 95% at the HQ USAREC level. Thus, the 80% PI for forecast data is ±1,389 contracts and the only departure from this band occurs in the very last month. Upon further visual inspection of Figure 18, the narrower (known inputs) 80% band of ±867 contracts seems equally useful. By contrast, the 95% PIs for both known and forecast inputs appear to be too wide for practical use.

Table 19. 80% and 95% Prediction Intervals by Contract Type, HQ USAREC-echelon

	80% PI		95% PI	
k	$\pm\hat{y}^{(k)}\|\mathbf{X}$	$\pm\hat{y}^{(k)}\|\hat{\mathbf{X}}$	$\pm\hat{y}^{(k)}\|\mathbf{X}$	$\pm\hat{y}^{(k)}\|\hat{\mathbf{X}}$
GA	436	686	667	1,049
OTH	291	461	445	705
SA	140	242	214	370
Total	**867**	**1,389**	**1,326**	**2,124**

Moving down to the contract level, Figure 19 provides the comprehensive MAPE and MAD for the estimation and validation sets, respectively. Again, the comprehensive comparison can be interpreted as per unit, per month. From the standpoint of percent error, SA contracts are predicted with more error than are GA and OTH; SAs also experience the greatest degradation in prediction accuracy for validation by far. However, the monthly production of SAs is also much smaller than either of the other two contract types. So in terms of absolute error, SAs have less deviation than do GA or OTH. This difference in interpretations of error between MAPE and MAD illustrates the need for both metrics.

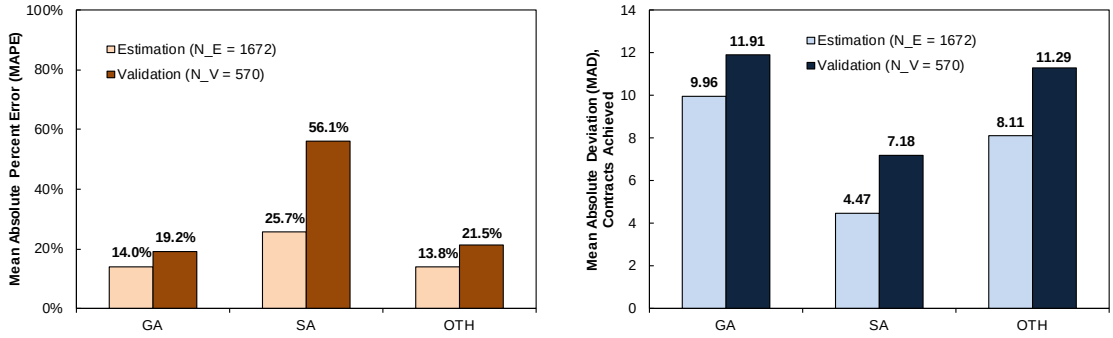


Figure 19. Comprehensive MAD and MAPE for the Three Contract Types

Since all contract types decreases in prediction accuracy between the estimation and validation data sets, we can determine whether or not these differences in performance are statistically significant. To address the question, we generated 95% confidence intervals—as shown in Table 20—on the MAD using the simple rule that the root mean squared 1-period forecast error, $\hat{\sigma} \cong 1.25 \cdot \text{MAD}$ [29]. Since there is

no overlap in the confidence intervals, we have reason to believe that the accuracy degradation—as measured by comprehensive MADs for each data set—is statistically significant for all three contract types. This result is consistent with our expectations although the relatively small magnitudes of performance degradation appear to indicate stable, predictive models.

Table 20. Mean Absolute Deviations (MAD) for the Three Contract Models with 95% Half-widths

k	$MAD^{(k)} \pm z_{0.05/2}\hat{\sigma}_E^{(k)}/\sqrt{N_E}$	$MAD^{(k)} \pm z_{0.05/2}\hat{\sigma}_V^{(k)}/\sqrt{N_V}$
GA	9.96 ± 0.60	11.91 ± 1.22
SA	4.47 ± 0.27	7.18 ± 0.74
OTH	8.11 ± 0.49	11.29 ± 1.16

For a closer look at each contract type, we provide the three respective HQ USAREC echelon plots in Figure 20. From Figure 20, the track and fit of GA and OTH are quite similar and in fact fairly strong, with R^2 equal to 0.70 and 0.73, respectively. In the GA estimation set we also note the recurrence of a failure to fit the latter 3 months of RY10, similar to what we observed for total contracts at the HQ USAREC echelon. Overall, the GA model tracks well until the last six months of validation where the convexity of the predictions lags behind the data. For the OTH model, we attribute its loss in accuracy to a sustained series of jumps in the actual data to around 2,500 contracts in the third quarter of RY14. This event appears to be a departure from previous behavior, especially considering the fact that GA historically tracked quite closely to OTH, although this jump is not simultaneously experienced by GA contracts. We wonder if this occurrence in the validation set represents a process departure similar to what we observed in RY10 for GA and total contracts. As with total contracts, 80% PIs with forecast inputs appear to be suitable for both GA and OTH contracts. For SA contracts, the seasonal high in the third quarter of RY14 is captured, albeit with less magnitude and perhaps a bit early. For some

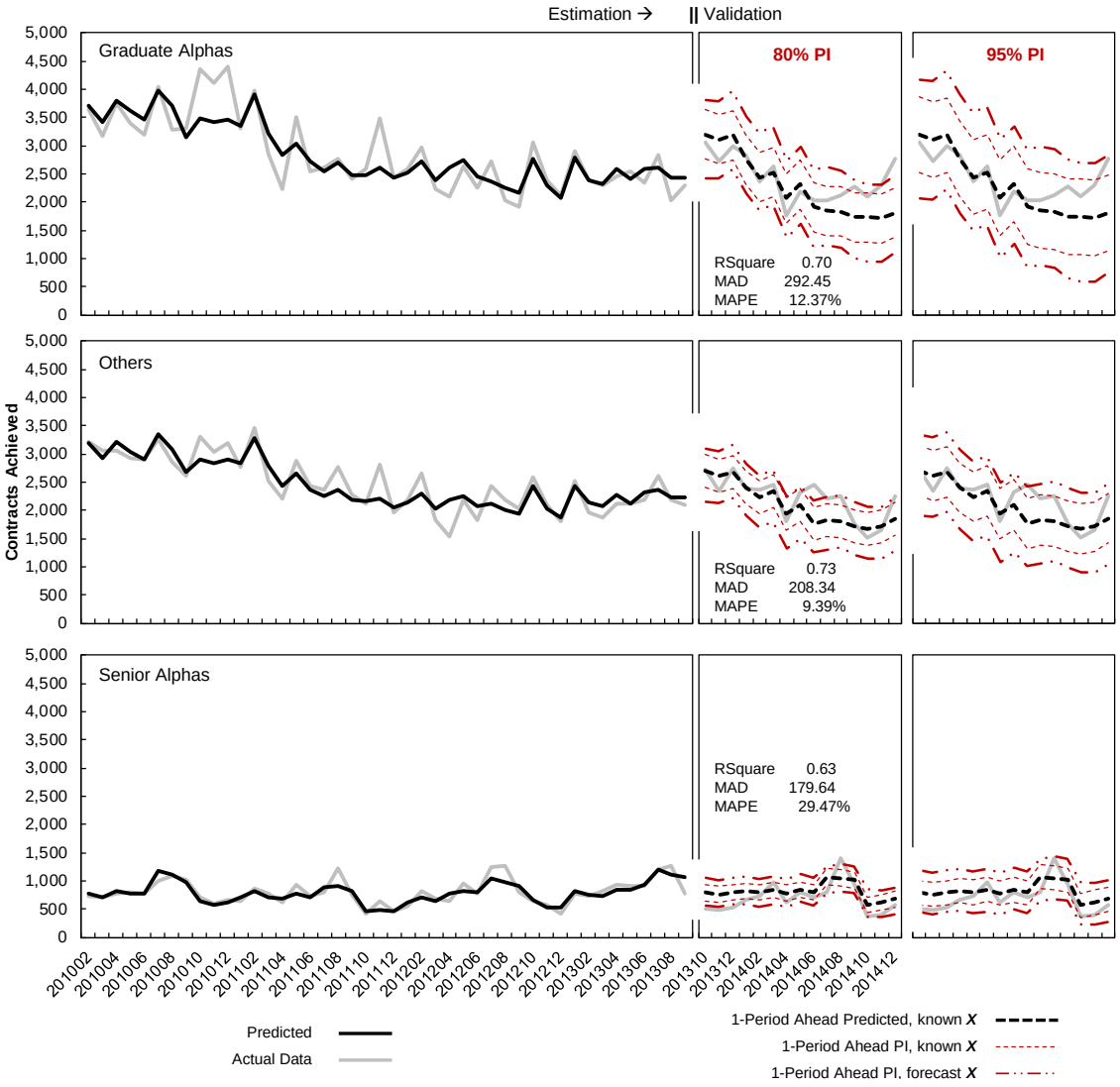


Figure 20. Time Series Data and Model Predictions for the Three Contract Types, USAREC Totals

reason, the seasonal fourth quarter trough is not captured nearly as well in RY13.

Overall, responses appear to be significantly more muted in the SA model during validation than in estimation; we are not sure why this occurred nor do we offer speculation. Nonetheless, all models appear to be approximating their respective processes with stable accuracy; this in itself is a promising indicator of their future utility, especially in light of previous results which had dramatically lower estimation fit metrics ($R^2 = 0.32$, 0.27 and 0.10 for GA, OTH, and SA respectively) and did not

address performance against a test set of data whatsoever [14].

Having analyzed validation performance from a macro-perspective, we finally delve down to the battalion echelon. It is fitting that we conclude our analysis here; insightful characterization of individual markets was our chief goal at the outset. In the proceeding pages, we present for each contract type a series of three charts which allow us to illustrate differences between the validation data and our models' battalion-level predictions. For each contract type, the left-hand charts depicts actual contracts achieved per month; it is sorted from top to bottom in decreasing order over the validation period. We provide the estimation data for visual reference but labels indicate validation values only. The middle chart gives decreases in MAPE between the estimation and validation sets. A dashed line represents the comprehensive accuracy degradation, which we use to discriminate above-average stability (to the right of the dashed line) from below-average stability (to the left of the dashed line) across the data-split. Finally, the right-hand chart shows each battalion's RMSE for both datasets while only labeling the validation values. The RMSE is useful in this case to provide an interpretation of ± 1 error standard deviation, assuming normality. However, we caution that such an assumption may not be valid for a battalion whose accuracy degradation is large.

The reading of these charts in a left-to-right sequence can be helpful in assessing both the potential production value and model accuracy for a given battalion market. Clearly the risks of assigning potential production value to a market with a poor model can be large. Furthermore, these risks are magnified at the extremes of observed contract production (i.e., highest- and lowest- producing battalions). So, it is perhaps useful to verify that battalions which perform at the extremes in also have small errors, or at least small degradations in errors between the estimation and validation data sets. For example, a model which appears to predict lots of contracts for a battalion

may be problematic if its accuracy is poor; in this manner blind faith in the model could lead to improperly high mission allocation. We examine these risks by assessing the top five and bottom five stable models for each contract type.

Following the discussions of each contract type by unit, we provide Figure 24 in the interest of added context. We have organized coropleth maps of the battalion areas which are color-coded by average actual contracts achieved per month over the validation period. We then summarize the the top five and bottom five battalions as described above. This method of display enables the reader to make a visual connection with the data, given the irregular unit boundaries that do not readily conform to standard geography. We recommend setting both Figure 24 and Appendix A aside for reference over the course of the next few paragraphs. Let us now begin by looking at GAs in Figure 21.

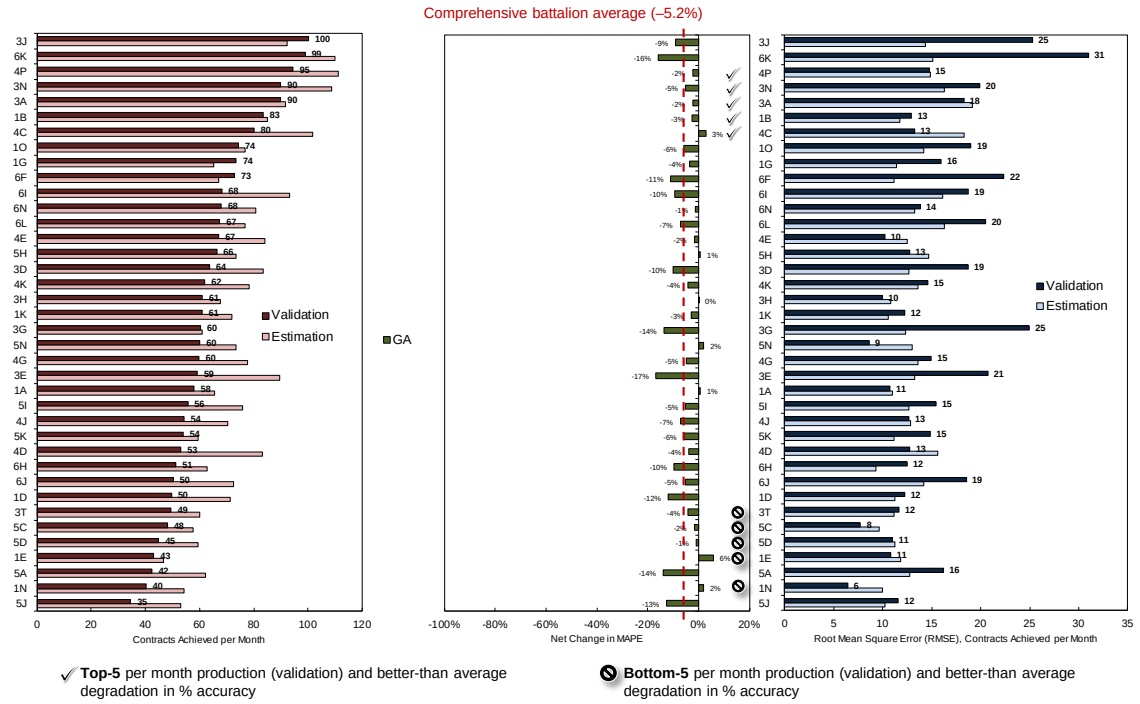


Figure 21. Model Performance with Estimation and Validation Data, GA by Battalion

GA Contracts. Battalions from 2nd BDE (i.e., BN RSIDs beginning with "3"), 5th BDE (i.e., RSIDs starting with "4"), and 6th BDE generally appear to produce more GA contracts in both datasets. However, we also note that several battalions from 2nd and 6th BDEs tend have greater than average degradation in prediction accuracy. As examples, we present the top two producing battalions, 3J (Raleigh) and 6K (Southern California); our model did not effectively predict the superior contract production in these two regions. We suspect this failure is due to a factor in these two markets which we did not include and by way of speculation, we observe that each has a relatively large military presence as indicated by the metric z_2, sponsor share. We ultimately omitted this variable from our model due to multicollinearity with other factors, although in these two regions it may have considerable impact. After BNs 3J and 6K, the five next highest-ranked battalions all experience very small prediction accuracy degradations as well as smaller RMSEs; this appears to indicate that our models in these markets are fairly accurate and we can therefore be confident in respective future predictions. Hence, BNs 4P (Phoenix), 3N (Tampa), 3A (Atlanta), 1B (Baltimore), and 4C (Dallas) are labeled in the "top five" of Figure 24.

Battalions from 1st BDE and 3rd BDE (i.e., RSIDs beginning with "1" and "5") dominate the lower end of the GA production spectrum. However, large error problems in BNs 5A (Chicago) and 5J (Milwaukee) prevent these markets from being adequately modeled. We do not offer speculation as to the content of these large errors, except that the urban environment in Chicago may account for a different type of error than the largely rural landscape of Wisconsin. By contrast, we are confident in the accurate prediction of low GA contract production in BNs 1N (Syracuse), 1E (Harrisburg), 5D (Columbus), 5C (Cleveland), and 3T (Baton Rouge). Battalions not listed in our top or bottom five either have mediocre production or errors too

large to consider the models as being accurate. Geographically, we note from Figure 24 that the dispersion of higher GA-producing markets is along the mid-Atlantic and Southwest regions; modeling accuracy is generally in agreement with the exception of the Tampa battalion. Under-performing GA markets are concentrated the Northeast and upper-Midwest, with which model accuracy also appears to generally agree. The notable exception is the Baton Rouge battalion, but we are not sure what factor may account for its increased model accuracy.

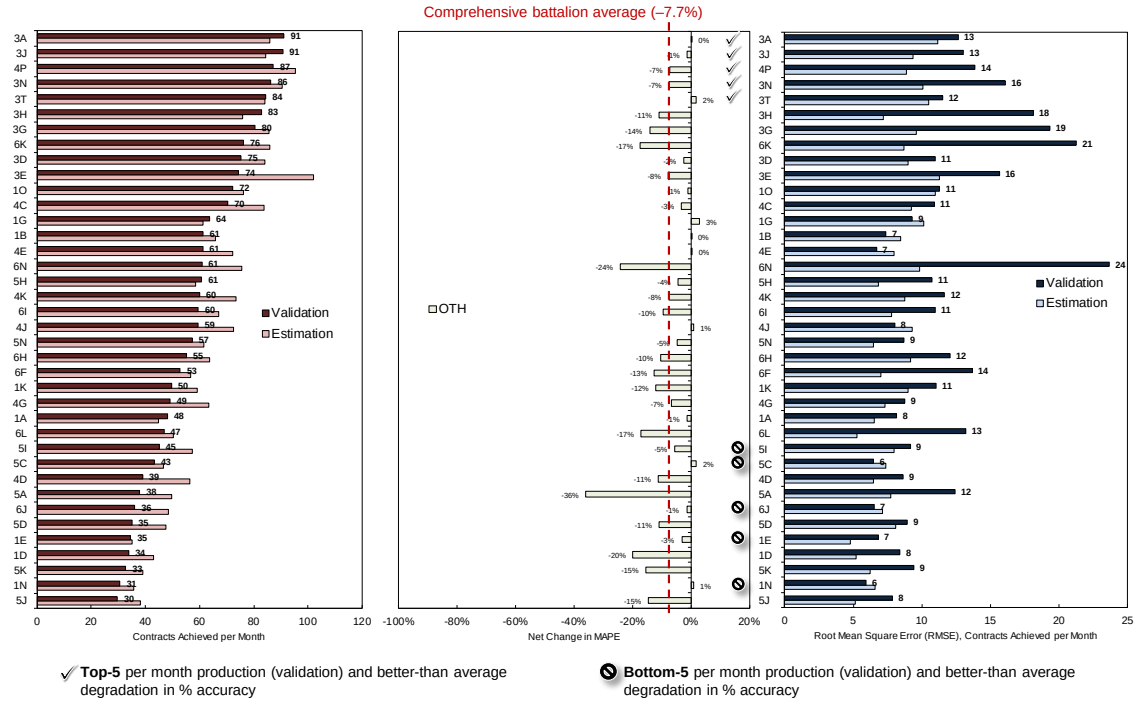


Figure 22. Model Performance with Estimation and Validation Data, OTH by Battalion

OTH Contracts. Interestingly, production of OTH contracts is geographically similar to that of GA contracts. We accurately capture top production in BNs 3A (Atlanta), 3J (Raleigh), 4P (Phoenix), 3N (Tampa), and 3T (Baton Rouge). So, the three markets in Atlanta, Tampa, and Phoenix have accurate models which predict top production of both contract types. In Baton Rouge, we see a reversal of production performance from GA to OTH, although accuracy remains consistently high.

And in Raleigh, which was a top performer but not accurately modeled as such with GA, we obtain both top production and accuracy with OTH contracts. This suggests we may have omitted a factor which was uniquely relevant to the high-aptitude high school graduate population in the Raleigh market. In general, we note both higher production and accuracy across the Southeast. Regarding OTH under-performers, the Mountain West constitutes a notable change from the GA geography. However, the dispersion of accurately modeled OTH under-performers is remarkably similar to that of their GA counterparts. Common to each category are BNs 1N (Syracuse), 1E (Harrisburg), and 5C (Cleveland). New additions are those of 5I (Great Lakes) and 6J (Salt Lake City). We hesitate to speculate on why the Mountain West region produces fewer OTH contracts or why some of the errors in the Northeast are large. However and regarding the former, cultural and religious factors in the densely populated Salt Lake City region may lend themselves to more high-quality contracts (less lower-quality).

SA Contracts. We recall that the SA model starts with a less accurate fit to the estimation data than GA or OTH contracts. Thus, we can expect more severe drops in performance against validation data at the market level. The average loss of 30% average percent accuracy appears to confirm this suspicion. However, we note that of the 38 battalion-level markets, only 12 have losses in accuracy greater than 30%; this indicates that perhaps the large average loss is being skewed by several outliers. As an example, we note that geographically contiguous BNs 3J (Raleigh), 3D (Columbia), and 3A (Atlanta) experience rather severe drops in percent accuracy ranging from 73% to 99%. Our speculation is limited, but we do offer again the possibility that high degrees of military presence in these markets may be impacting production. The installations of Fort Bragg, Fort Jackson, and Fort Gordon are located in this region and we conceive that high school senior populations may be

more sensitive to changes in local military environments. This explanation may have additional plausibility with regard to BN 6K (Southern California), which has the highest largest military presence of all markets. However, this explanation is not satisfactory for other high-error units with low military presence such as 3G (Miami) or 5A (Chicago).

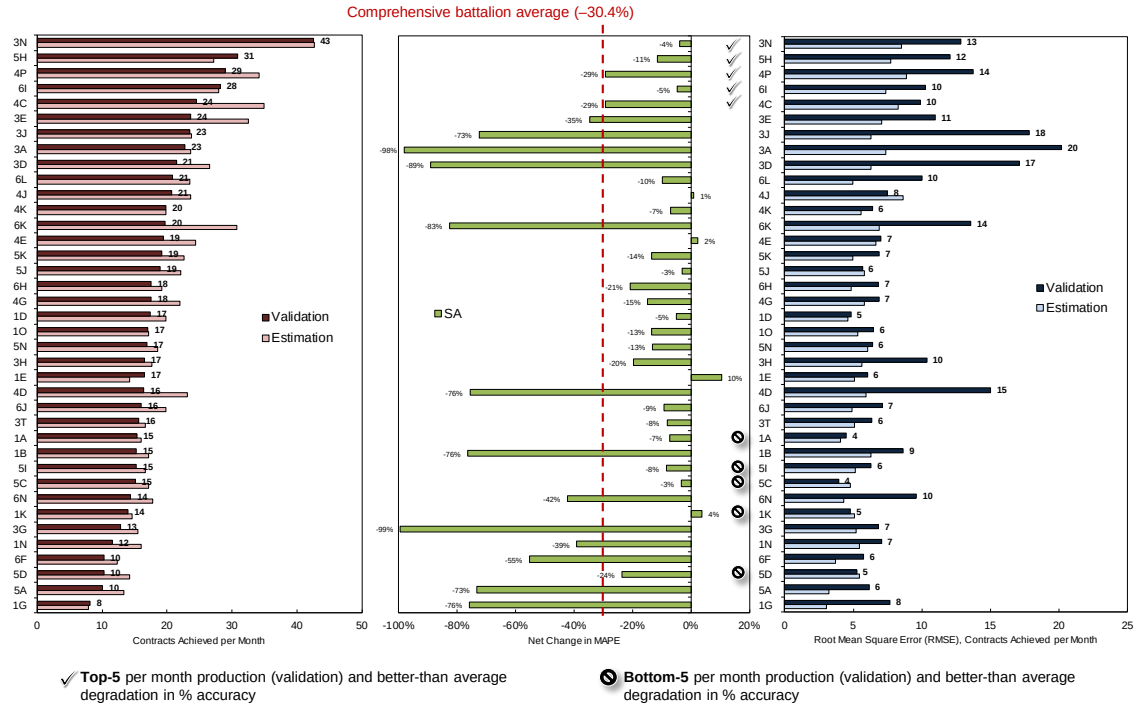


Figure 23. Model Performance with Estimation and Validation Data, SA by Battalion

Now regarding the extreme performers: the five highest-producing units are also accurately modeled. Interestingly, BN 3N (Tampa) is the most accurately predicted and by far the highest SA-producing of all battalions during both estimation and validation; we placed this unit in the top five for GA and OTH as well. Battalion 4P (Phoenix) also has a top-five ranking for all three contract types. Battalion 4C (Dallas) is ranked high for both SA and GA contracts. Battalions 5H (Indianapolis) and 6I (Sacramento) round out the top five accurate SA contract models; these areas are relatively unique, having not figured prominently in either GA or OTH contract

models. We are not exactly sure why this is the case since there does not appear to a readily intuitive similarity between these regions. For bottom-performing, yet accurate SA models we once again turn to the Northeast. Battalions 5C (Cleveland), 5D (Columbus), and 5I (Great Lakes) are recurring entires. Rounding out the bottom-five SA models are BNs 1A (Albany) and 1K (Mid-Atlantic), which are also new entires. Upon inspection of Figure 23, the choice of the SA bottom-five is made more difficult due to frequent occurrences of large accuracy losses. In each of the problematic markets (e.g., BNs 1B, 3G, etc.) the model predicted significantly more contracts than were actually produced; this indicates some factor is not being properly accounted for in these areas. Unfortunately, there does not appear to be a common factor linking these areas so we do not offer speculation as to its nature.

Closing Remarks. We are now at the end of our analysis, and are nearing completion of our overall task at-hand. By way of transition, let us concisely recapitulate our results and introduce several assertions to the basis for our concluding chapter. Herein we have described our process of constructing linear regression models using a unique set of variables that characterize specific recruiting markets and contract types. Given the consistently accurate predictions of these models in the face of new data, we have shown that all three models possess some degree of future utility. Relative to the RMI currently in use, our models convey more information regarding the response; moreover, we utilize a set of predictor variables that is more independent and arguably more universally descriptive. Also, the RMI models two contract types—GA and SA—together in the same response. Yet our efforts have shown that these two contracts respond differently in nearly every respect. In and of itself, this is a valuable insight. Finally, the accuracies of our models as measured by R^2 far exceed those in previous studies. Yet as always, there is room for improvement; we will address this fact in due course in our next and final chapter.

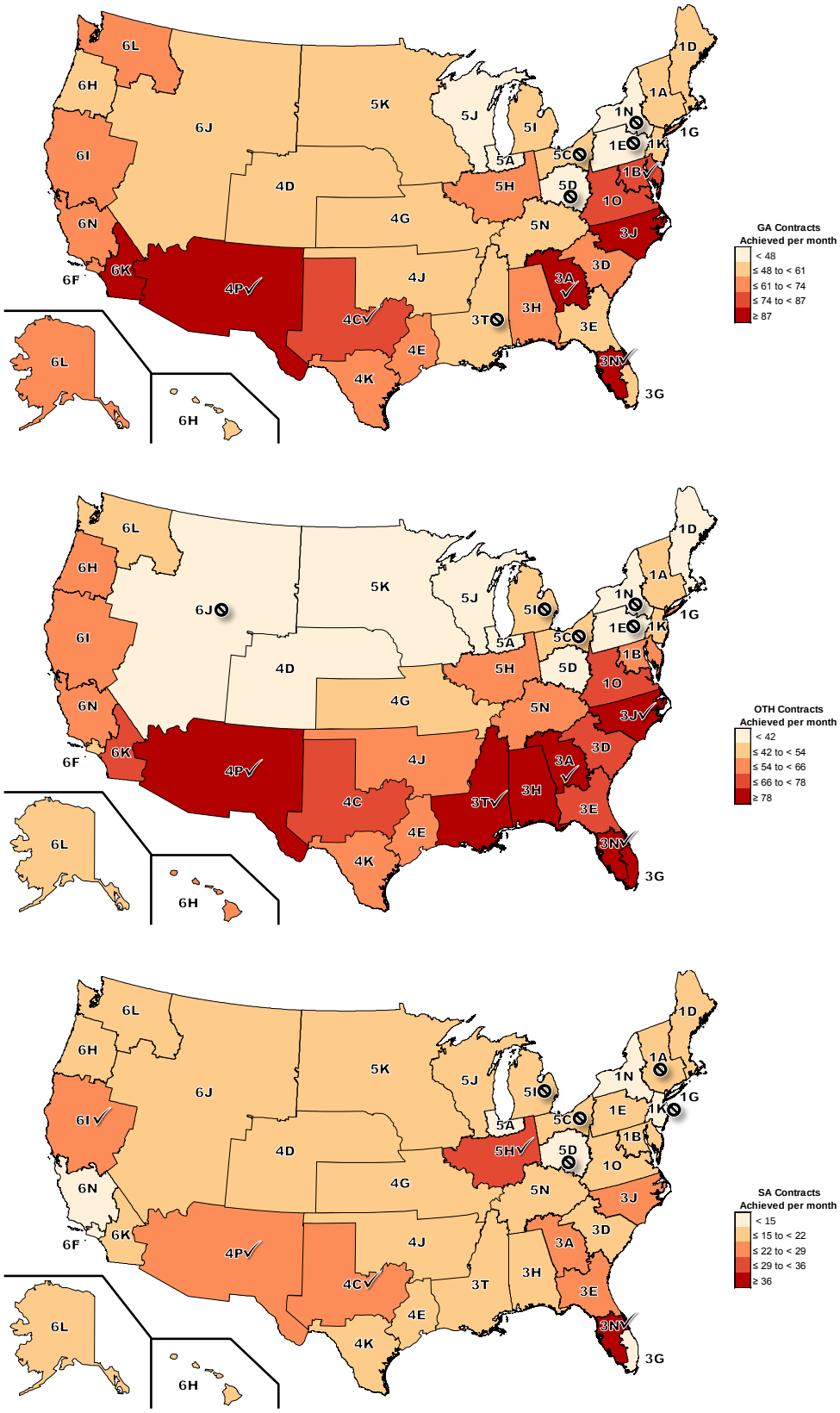


Figure 24. Choropleth Map of Contracts Achieved per Month (Validation Data Only), with Top-five and Bottom-five Battalion Models

V. Conclusion

As we begin our closing chapter, we are mindful of the contributions of our research with respect to three key areas. First, we wish to provide an assessment of how closely our results agree with the findings of previous literature, which we covered in Chapter II. Second, we need to summarize the utility of our approach as compared with current USAREC procedures. Lastly and prior to our closing remarks, we we take an honest look at the limitations of our research and suggest several ways upon which it could be improved.

5.1 Comparisons with Previous Literature

Making quantitative comparisons between our research and previous literature is somewhat challenging for a couple of reasons. The first reason encompasses differences between the definition and reporting of independent parameters in the models. Much of the literature we reviewed contained transformed independent variables, whereas our modeling approach did not necessitate such transformations. Therefore, a direct comparison of magnitudes between independent terms common to our and previous research (e.g., unemployment) is not possible and we refrain from attempting it. We can, however, provide a general assessment regarding the relative importance for at least two common independent variables: recruiting missions and unemployment.

All three of our models show that the mission is more important, generally, than unemployment.[1] While previous studies are not in universal agreement over the exact magnitudes of these two effects, there appears to be consensus regarding the relative importance we just presented, as well as the fact that both are positively correlated with contracts achieved. The one exception was our finding that unemployment

[1]Recall that for individual units, this may differ although the majority do not deviate from the magnitudes of each main effect. See Appendix F for individual unit models details.

was negatively correlated with SA contracts achieved, although as we mentioned in Chapter IV this may be due to a coincidental cycle between unadjusted unemployment and SA contracts that is not present with GA or OTH contracts. Unfortunately, substantial differences in between the remainder of independent variable specifications preclude further comparisons of such effects.

A second difficulty in the comparison of results is related to the first, but instead deals with differences in dependent variable definitions. In fact, only one study—that of Dertouzos and Garber (2008)—is directly comparable for the three responses of GA, OTH, and SA contracts that we used in our research. In the former study the authors achieved R^2 values of 0.32, 0.27 and 0.10 for GA, OTH and SA contracts, respectively [14] when using the month as the time unit of observation; by contrast, we showed that for validation data our models achieved R^2 values equal to 0.70, 0.73 and 0.63 for the same respective contract types. Using our validation fits as a baseline, our results still provide considerable relative improvements of 530%, 170%, and 119% for SA, OTH, and GA contracts, respectively. Figure 25 presents the R^2 values for our research and that of Dertouzos and Garber (2008). We denote our baseline for the primary relative comparison with an asterisk.

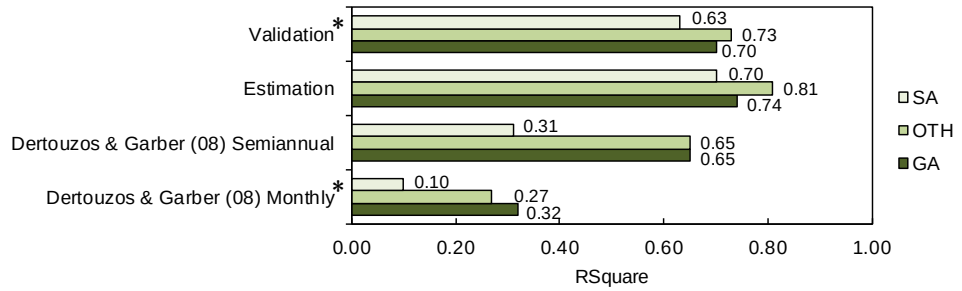


Figure 25. Model Fits From This Research Compared with Previous Literature

The improvements are even more pronounced when we examine our metrics of fit for the estimation data, for which we obtained R^2_{Adj} of 0.74, 0.81 and 0.70, respectively (non-adjusted R^2 are higher by about 0.01 in each case). Now when using a six-month

aggregate unit of observation, Dertouzos and Garber were able to achieve R^2 of 0.65 for GA and OTH and 0.31 for SA [14]. Notwithstanding this improvement by the previous study, our model fit results remain superior while maintaining the use of monthly data.

We do finally note a dramatic difference in the number of independent variables we used as compared with previous literature. We used a set of five continuous variables compared to previous studies (from Appendix B) which used 21 continuous independent variables on average. We do sacrifice some of the apparent gains in parsimony by including between 37 and 40 indicator variables for battalions and quarters, but these additions are necessary to achieve market specificity with respect to each contract model. In fact, we note that the inclusion of market-specific models—which we identified early on as a key attribute of decision-maker utility—is another key improvement over *all* previous efforts. Thus, we have shown that a refined set of variables can produce fairly superior models, both in terms of fit to the data and utility to decision-makers.

5.2 Comparison to Current USAREC Models

The current method of missioning at USAREC incorporates a three-step series of weighting numerous contract achievement factors by ZIP codes. The RMI, which is the subject of our initial analysis in Chapter IV, is only the second step in this weighting process; as such, its outputs are not meant to produce direct predictions of contracts achieved as it is not the final step in the missioning process. Our models, on the other hand, do produce forecasts of contract production in a single step. This fact alone could present a valid argument for implementation of our models.

Nevertheless, we have some concern that the current specification of the RMI itself is problematic. For one, it contains significant multicollinearity—between the

required quality per recruiter and required volume per recruiter terms—which must be addressed. However, the greater issue lies with the inclusion of a dependent variable, achieved volume per recruiter, in the independent variable set. The fit of this model—which has $R^2 = 0.93$—appears to be so good as to render subsequent improvement efforts futile. However, a truly correct specification requires removal of the volume per recruiter term, or its inclusion in lagged form. For simplicity we chose as the baseline the form of the RMI without this term entirely. Thus, when the RMI is specified with a correct set of independent variables, we see an R^2 of 0.70 which much more closely resembles the results we obtained. Given that all three of our contract models obtained adjusted correlation coefficients greater than or equal to the unadjusted correlation of the RMI, it appears that collectively they provide a modest improvement over the RMI assuming its specification less volume per recruiter in an independent role.

Perhaps the most salient observation we can make in comparison to the RMI (fit metrics aside) is our distinction between GA and SA contracts. The RMI includes GA and SA contracts achieved per recruiter in the same ratio, implying the two contracts have similar responsive attributes. However, we observed beginning with the literature review and continuing with variance reduction and ultimately model fits, that these two contracts respond much differently from one another. Given these differences, it would seem counter-productive to continue modeling these contract types simultaneously within a single response. Furthermore, the relative strength of our models should be an encouraging sign that separating the two contract types can be undertaken without significant loss of accuracy.

5.3 Future Research Opportunities

Even with the promising results we obtained, there is ample opportunity for future exploration. First, we note that the list of potential regressors can always be expanded. Two key factors come to mind, which are those of fiscal expenditures Army RC missions; previous literature indicated both of these to be important to one degree or another. In the case of expenditures—which can be further categorized by normal operations and support or advertising costs—the data was simply not available for enough of the estimation data. We attempted to include competition from RC missions indirectly by virtue of recruiter share, but this term was discarded during the variance reduction stage of our analysis due to multicollinearity. Perhaps we should have included separate RC missions themselves, similar to the active duty missions which proved to be highly influential, and this is a lesson learned.

Third, and perhaps most importantly, the utility of our models may not be fully realized by HQ USAREC until the models of each contract type can be unified into a single, comprehensive, deterministic platform. By this we mean that the controllable regression model variables (i.e., GA+OTH mission, Req_SA_PR) for all three contract types be simultaneously decided upon in such a way so as to maximize predicted contracts subject to organizational constraints. As examples of organizational constraints for USAREC, we suggest the high-school graduate and quality requirements imposed by Congress and/or inter-battalion balances of mission equity. An additional constraint should also prevent extrapolation beyond the input data used to estimate the models to ensure that the coefficients remain properly predictive. And while a single objective such as contract maximization may be useful, we suspect that added utility may be present with the incorporation of goals such as minimization of prediction error. The unit-specific measures of prediction accuracy we included at the conclusion of Chapter IV could easily be leveraged for this purpose, thereby ensuring

that risk is sufficiently accounted for in the allocation of recruiting resources. If it can be shown that a USAREC-level contracts achieved prediction interval—obtained from any type of optimization routine such as the two we have just described—differs significantly from the PI obtained using actual inputs, a valid case for the utility of the former may exist.

5.4 Final Remarks

We have shown that adequate models for each of the three major Army Active Duty contract types—high-aptitude high school graduates, high-aptitude high school seniors and others—can be quantitatively estimated with individual accuracies greater than or equal to both previous literature as well as the current model in use by HQ USAREC. Furthermore, we have shown that such models can be parsimoniously estimated according to the required assumptions of model adequacy. We have then shown that all three models perform well in the face of new data, achieving close to two-thirds prediction accuracy in the worst case of high-aptitude high school seniors. Finally, we note that all these results were achieved using largely open-source data, which was innovatively mapped to battalion-specific areas using county-level data. In light of these results, we make the case for implementation of our modeling approach for Army recruiting—and indeed for any private sector marketing strategy—as well as for continuing efforts to achieve greater granularity with respect to geographical data. Although predicting the future is inherently difficult, our attempt at doing so appears to light an encouraging path forward.

Appendix A. Unit Recruiting Station Identifications (RSID)

Table A.1. Brigade RSIDs

RSID	Unit - Headquarters (Region)
1	1st BDE - Ft. Meade (Northeast / Mid-Atlantic)
3	2nd BDE - Redstone Arsenal (Southeast)
4	5th BDE - Ft. Sam Houston (Plains)
5	3rd BDE - Ft. Knox (Upper Midwest)
6	6th BDE - Las Vegas (Mountain / West)

Table A.2. Battalion RSIDs

RSID	Region
1A	Albany
1B	Baltimore
1D	New England
1E	Harrisburg
1G	New York City
1K	Mid-Atlantic
1N	Syracuse
1O	Richmond
3A	Atlanta
3D	Columbia
3E	Jacksonville
3G	Miami
3H	Montgomery
3J	Raleigh
3N	Tampa
3T	Baton Rouge
5A	Chicago
5C	Cleveland
5D	Columbus
5H	Indianapolis
5I	Great Lakes
5J	Milwaukee
5K	Minneapolis
5N	Nashville
4C	Dallas
4D	Denver
4E	Houston
4G	Kansas City
4J	Oklahoma City
4K	San Antonio
4P	Phoenix
6F	Los Angeles
6H	Portland
6I	Sacramento
6J	Salt Lake City
6K	Southern California
6L	Seattle
6N	Fresno

Appendix B. Variables Used in Past Studies

Table B.1. Dependent Variables in Reviewed Literature

Variable	Dertouzos (1985)	Kilburn, et al ('99)	Murray, et al ('99)	Warner, et al ('01)	Kleykamp (2006)	Dertouzos, et al ('06)	Dertouzos, et al ('08)	Asch, et al ('09)	Gibson, et al ('09,'11)
Contracts, high-quality men							y		
Contracts, high-quality women							y		
Contracts, other men							y		
Contracts, other women							y		
Contracts, high-quality graduates			y				y		
Contracts, high-quality seniors			y				y		
Contracts, high quality White graduate								y	
Contracts, high quality African-American graduates								y	
Contracts, high quality Hispanic graduates								y	
Contracts per Recruiter, high-quality						y			
Contracts per Recruiter, All						y			
Contracts per Station, high quality						y			
Number of accessions									y
Probability of choosing to enlist		y			y				
Probability of choosing to attend college		y			y				
Probability of choosing to work/other		y			y				
Enlistments, other graduates	y		y^*						
Enlistments, high quality graduates	y		y^*	y					
Propensity				y					
Probability of DEP Attrition				y					

*lagged

Table B.2. Independent Variables in Reviewed Literature: Advertising & Demographic

Variable	Broad Category	Dertouzos (1985)	†Kilburn, et al ('99)	Murray, et al ('99)	†Warner, et al ('01)	†Kleykamp (2006)	Dertouzos, et al ('06)	Dertouzos, et al ('08)	Asch, et al (2009)	Gibson, et al ('09,'11)
Advertising spending by MUD-Navy	Advertising				x					x
Total Army advertisements in last 11 months	Advertising				x					
Total Army television ads in last 11 months	Advertising				x					
Total Army non-T.V. ads in last 11 months	Advertising				x					
Total joint ads in last 11 months	Advertising				x					
Total joint T.V. ads in last 11 months	Advertising				x					
Total joint non-T.V. ads in last 11 months	Advertising				x					
Ratio of QMA population to OPRA recruiters	Demographic						x	x		
Ratio of African American men to total men	Demographic						x	x		
Ratio of Hispanic men to total men	Demographic						x	x		
Percentage of 17–21 year old male population in college	Demographic							x		
Ratio of urban population ($\geq 50,000$ per US Census) to total population	Demographic							x		
Ratio of urban cluster population $(2,500 \leq p < 50,000)$ to total population	Demographic							x		
Ratio of single-parent households in year 2000 to year 1990	Demographic							x		
Ratio of children in poverty to total population	Demographic							x		
Ratio of professed adult Catholics to total propulation	Demographic							x		
Ratio of adults professing an Eastern religion to total propulation	Demographic							x		
Ratio of professed non-Catholic Christians to total propulation	Demographic							x		
Ratio of veteran population, age ≤ 32, to young male population (age 17–21)	Demographic						x	x		
Ratio of veteran population, age 33–42, to young male population (age 17–21)	Demographic						x	x		
Ratio of veteran population, age 43–55, to young male population (age 17–21)	Demographic						x	x		
Ratio of veteran population, age 56–65, to young male population (age 17–21)	Demographic						x	x		
Ratio of veteran population, age 65–72, to young male population (age 17–21)	Demographic						x	x		
Ratio of veteran population, age ≥ 73, to young male population (age 17–21)	Demographic						x	x		
Ratio of recruiters to size of the adult population	Demographic					x			x	
Percent veteran	Demographic				x				x	x
Percent non-citizen	Demographic								x	
Percent obese	Demographic								x	
Percent college enrollment	Demographic				x		x		x	
Average age	Demographic									x
Correctional facility population	Demographic									x
Population density	Demographic				x					x
Service members	Demographic									x
Military casualty count	Demographic									x
Percent African-American, high school population	Demographic									x
Percent Hispanic, high school population	Demographic									x
Proportion of population with asthma	Demographic									x
High school population	Demographic									x
Percent county employment from military	Demographic					x				
Percent African-American	Demographic		x		x	x				
Percent Hispanic	Demographic		x		x	x				
Percent female	Demographic					x				
Population of males aged 15–19 years	Demographic									
Percent of labor force female	Demographic		x							
Percent of the male population male and aged 18–24	Demographic		x							
High quality African-Americans available	Demographic									
High quality Hispanic available	Demographic									
High quality available (all)	Demographic									
Percent QMA	Demographic				x					

†Variables obtained from individual survey data are not shown.

Table B.3. Independent Variables in Reviewed Literature: Geographic, Mission, Political, & Recruiter

Variable	Broad Category	Dertouzos (1985)	†Kilburn, et al ('99)	Murray, et al ('99)	†Warner, et al ('01)	†Kleykamp (2006)	Dertouzos, et al ('06)	Dertouzos, et al ('08)	Asch, et al (2009)	Gibson, et al ('09,'11)
Mountain (binary)	Geographic						x	x		
North Central (binary)	Geographic						x	x		
South (binary)	Geographic						x	x		x
Pacific (binary)	Geographic						x	x		
Average July temperature	Geographic							x		
July average precipitation	Geographic							x		
July average humidity	Geographic							x		
Northeast region	Geographic									x
West region	Geographic									x
Distance to nearest college or university	Geographic									x
Distance to nearest military installation	Geographic									x
Size of nearest college or university	Geographic									x
Size of nearest military installation (in 10,000 personnel)	Geographic									x
Distance to nearest Air Force recruiting office	Geographic									x
Distance to nearest Coast Guard recruiting office	Geographic									x
Distance to nearest college or university-squared	Geographic									x
Distance to nearest Marine Corps recruiting office	Geographic									x
Distance to nearest Navy recruiting office	Geographic									x
Mission, high-quality seniors plus DEP loss	Mission							x		
Mission, high-quality graduates plus DEP loss	Mission							x		
Mission, others plus DEP loss	Mission							x		
Percent of national enlistments, combat support MOSs	Mission							x		
Percent of national enlistments, white-collar MOSs	Mission							x		
Percent of national enlistments, blue-collar MOSs	Mission							x		
Percent of national enlistments, combat MOSs	Mission							x		
Ratio of SA production to prev. year mission with 3-month lag	Mission							x		
Ratio of GA production to prev. year mission with 3-month lag	Mission							x		
Ratio of OTH production to prev. year mission with 3-month lag	Mission							x		
Recruiter Goal, high quality (mission plus DEP losses)	Mission	x		x	x		x			
Recruiter Mission, high quality (excluding DEP losses)	Mission						x			
Station Mission, high quality (excluding DEP losses)	Mission						x			
Iraq War effect	Political								x	
President Bush approval rating	Political								x	
RA contracts as percentage of total DoD active duty contracts, 1999	Production						x	x		
Mission- Air Force level 2 MUD meeting mission	Production									x
Mission- number of Marine recruiting offices that made mission	Production									x
Number of Regular Army Recruiters on production	Recruiter							x		
2-Recruiter Station (binary)	Recruiter						x	x		
3-Recruiter Station (binary)	Recruiter						x	x		
4-Recruiter Station (binary)	Recruiter						x	x		
5-Recruiter Station (binary)	Recruiter						x	x		
≥6-Recruiter Station (binary)	Recruiter							x		
Ratio of on-production commander to on-production recruiters	Recruiter							x		
Ratio of (non-production) recruiters on duty to on-production recruiters	Recruiter							x		
Ratio of (non-production) recruiters absent to on-production recruiters	Recruiter							x		
Ratio of (non-production) commanders to on-production recruiters	Recruiter							x		
Recruiter demographics (20 various incl. race, ed cat, AFQT cat, MOS, etc.)	Recruiter						x			
Recruiters-Army	Recruiter	x		x	x					x
ASVAB tests given in high schools	Recruiter									x

†Variables obtained from individual survey data are not shown.

Table B.4. Independent Variables in Reviewed Literature: Reserve/Joint, Resource, Socio-economic, & Time

Variable	Broad Category	Dertouzos (1985)	†Kilburn, et al ('99)	Murray, et al ('99)	†Warner, et al ('01)	†Kleykamp (2006)	Dertouzos, et al ('06)	Dertouzos, et al ('08)	Asch, et al (2009)	Gibson, et al ('09,'11)
Ratio of RC recruiters to OPRA recruiters	Reserve, Joint							x		
Ratio of RC"OTH" mission to number of OPRA recruiters	Reserve, Joint							x		
Ratio of RC prior service mission to number of OPRA recruiters	Reserve, Joint							x		
Ratio of RC high school mission to number of OPRA recruiters	Reserve, Joint							x		
Ratio of RC "OTH" DEP loss to number of OPRA recruiters	Reserve, Joint							x		
Ratio of RC prior DEP loss to number of OPRA recruiters	Reserve, Joint							x		
Ratio of RC high school DEP loss to number of OPRA recruiters	Reserve, Joint							x		
Recruiters-Air Force	Reserve, Joint									x
Recruiters-Army Guard	Reserve, Joint									x
Recruiters-Army Reserve	Reserve, Joint									x
Recruiters-Coast Guard	Reserve, Joint									x
Recruiters-Marine Corps	Reserve, Joint									x
Recruiters-Navy	Reserve, Joint									x
Bonus accessions- Marine Corps	Reserve, Joint									x
Bonus accessions-Air Force	Reserve, Joint									x
Total sister-service mission, high quality	Reserve, Joint				x					
Enlistment bonus, average total offered in cash	Resource				x				x	
Ratio of national maximum MGIB benefit to average state college tuition	Resource								x	
Percentage of new recruits offered the Army College Fund	Resource				x				x	
Enlistment incentives- Navy, cash only	Resource									x
Enlistment incentives- Navy, total	Resource									x
Average business size (in employees)	Socio-economic									x
Average vehicles per household	Socio-economic									x
Change in unemployment from previous month	Socio-economic							x		
College entrance test-ACT composite scores	Socio-economic									x
Crime rate	Socio-economic								x	
English proficiency	Socio-economic									x
Government workers	Socio-economic									x
Household effective buying income (in hundreds of thousands of $)	Socio-economic				x					x
Households with no vehicles	Socio-economic									x
Per capita income	Socio-economic					x				
Percent change in per capita personal income	Socio-economic		x							
Population in poverty	Socio-economic									x
Property crimes	Socio-economic									x
Proportion of college graduates	Socio-economic									x
Proportion of population married	Socio-economic									x
Proportion of population smoking every day	Socio-economic									x
Ratio of manufacturing earnings to E-4 monthly salary	Socio-economic						x	x		
Ratio of military to civilian wages	Socio-economic			x	x				x	
SAT scores-math	Socio-economic									x
Subject test-algebra scores	Socio-economic									x
Unemployed	Socio-economic			x						
Unemployment rate	Socio-economic	x				x	x	x	x	x
Unionized workers	Socio-economic									x
Violent crimes	Socio-economic									x
Volunteers	Socio-economic									x
Wages for manufacturing production workers	Socio-economic	x								
Weighted average tuition	Socio-economic									x
January (binary)	Time				x		x			
February (binary)	Time				x		x	x		
March (binary)	Time				x		x	x		
April (binary)	Time				x		x	x		
May (binary)	Time				x		x	x		
June (binary)	Time				x		x	x		
July (binary)	Time				x		x	x		
August (binary)	Time				x		x	x		
September (binary)	Time							x		
October (binary)	Time				x		x	x		
November (binary)	Time				x		x	x		
December (binary)	Time				x		x	x		

†Variables obtained from individual survey data are not shown.

Appendix C. ZIP Code Crosswalk Procedure

As we mentioned in Chapter III, the crosswalk between ZIP codes and counties required a multi-step procedure. This stems from the fact that no reliable means of directly linking counties to directly ZIP codes exists, as we describe subsequently. However, we do show that ZIP codes can be effectively linked similar geographies known as ZIP Code Tabulation Areas (ZCTAs). After performing this intermediate step, ZCTAs can then be mapped directly to counties which completes the desired original linkage. This Appendix covers our exploratory investigation and ultimate resolution of these mapping processes.

We begin our discussion at the most basic mapping level, which involved matching every ZIP code with its respective recruiting unit. This step was relatively easy since for each echelon (i.e., brigade, battalion, company, center) recruiting units are defined by a mutually exclusive set of ZIP codes. We initially matched ZIP codes to the center level but learned this would be too difficult to track in past years since lower-echelon boundaries change much more frequently than those of battalions or brigades. Therefore, we matched only to the battalion and brigade echelons as indicated by the following pseudo-code:[1].

```
FOR Each ZIP code In z5max.xlsx
   IF ZIP code is not in 50 states or D.C. THEN Remove record
   ELSE
      FOR each ZIP and FIPS < 5 chr.
         Add leading ZIP zeros (CT,MA,ME,NJ,NY/Fishers Is.,RI,VT)
         Add FIPS zeros (AL-CT)
         Format as text
      NEXT ZIP-FIPS string
      FOR each USAREC echelon (BDE,BN,CO,CTR)
            Assign Echelon to ZIP Code with ZIPs_by_RSID.xlsx
         IF echelon is not found THEN assign closest contiguous CTR--BDE
      NEXT Echelon
   END IF
NEXT ZIP code
```

[1]FIPS is an acronym that stands for Federal Information Processing Standards; FIPS are 2-digit and 3-digit numerical codes that indicate states and counties, respectively. The original crosswalk file given by USAREC ("z5max.xlsx") required some cleaning to convert FIPS codes to a uniform, usable format.

Next, we located the master files that correlate ZIP codes to counties by a variety of metrics. These files are available from the Department of Housing and Urban Development (HUD) [47]. We attempted to match the two files with the following approach:

```
Revised ZIPs + units = ZIP_CountyFIPS_AOs.xlsm
FOR 2QTRFY15,1QTRFY15,1QTRFY14,1QTRFY13,1QTRFY12,1QTRFY11,1QTRFY10
   Download HUD ZIP-to-County Correlations from
      http://www.huduser.org/portal/datasets/usps_crosswalk.html
      FOR Each ZIP-FIPS pair (1 to 40,999) IN ZIP_CountyFIPS_AOs.xlsm
         FOR Each HUD correlation file (1 to 7)
            Assign percent of total ZIP addresses in county (FIPS)
         NEXT HUD file
      NEXT ZIP-FIPS Pair
   Save correlation file as hidden tab to ZIP_CountyFIPS_AOs.xlsm
NEXT QTR
```

At this point, we noted poor accuracy in the matching process. Table C.1 shows this problem being exacerbated with each previous year. We began to assess alternatives to matching ZIP codes directly. ZCTAs are analogous to ZIP codes; the former is used by the Census Bureau while the latter is strictly a US Postal Service construct for delivery routes. Also, ZCTA boundaries are likely to be more constant over time [2]. However, the Census Bureau does not publish a direct correlation file between ZIP codes and ZCTAs [48]. It does, however, give ZCTA-county correlations for Census 2010. A graphical summary of the difference between ZCTAs and ZIP codes is in Figure C.1.

Table C.1. Accuracy of Housing and Urban Development (HUD) ZIP Code-to-County Correlation Files

	2QTR15	1QTR15	1QTR14	1QTR13	1QTR12	1QTR11	1QTR10
Number ZIP codes Unmtached	1883	1893	2263	2278	2378	5050	5149
As % of Total ZIPs	4.6	4.6	5.5	5.6	5.8	12.3	12.6

A public database known as "UDS Mapper" provides a ZIP code to ZCTA crosswalk for calendar year (CY) 2014 [49]. UDS Mapper is a joint venture between the Department of Health and Human Services (DHHS) and the Robert Graham Center, a body of clinical researchers, social scientists, economists, and geographers. UDS

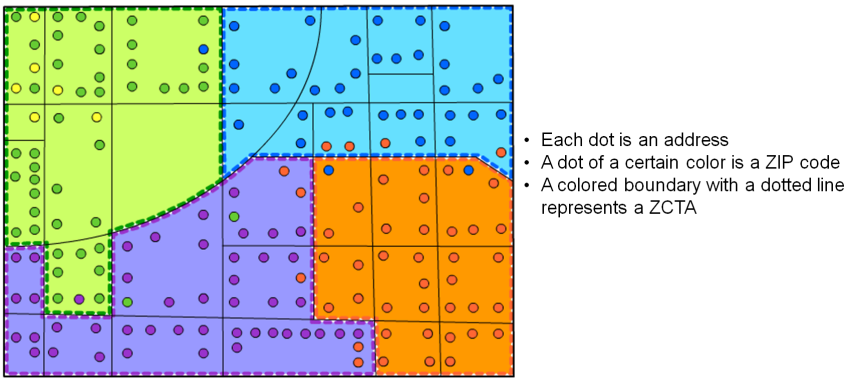


Figure C.1. Overview of ZCTA Design (Source: U.S. Census Bureau [2])

stands for the Uniform Data System [50]. From the website, "[DHHS], John Snow, Inc. and the Robert Graham Center have collaborated to develop a mapping and decision-support tool driven primarily from data within the Uniform Data System (UDS), previously not publicly accessible at the local level [50]."

After downloading this file we repeated the matching procedure in a similar manner. However, the ZCTAs allowed us to then use Census Bureau files to match county populations, as is shown by the following example code [27, 28].

```
WITH CY2014 (most current available)
   Download the ZIP-code to ZCTA crosswalk file available from
      http://www.udsmapper.org/zcta-crosswalk.cfm
   Format fields as text (i.e., retain leading zeros)
   FOR Each ZIP code (1 to 40,999) In ZIP_CountyFIPS_AOs.xlsm
      Assign ZCTA
   NEXT ZIP code
   Save crosswalk file as tab in ZIP_CountyFIPS_AOs.xlsm
END WITH

WITH 2010 Census
   Download ZCTA to County correlation file available from
      https://www.census.gov/geo/maps-data/data/zcta_rel_download.html
   Format fields as text (i.e., retain leading zeros)
   FOR Each ZIP-code (1 to 40,999) In ZIP_CountyFIPS_AOs.xlsm
      Assign percent ZCTA population residing in applicable county(counties)
   NEXT
   Save correlation file as tab in ZIP_CountyFIPS_AOs.xlsm
END WITH
```

At this point, all but 15 ZIP codes were successfully matched to ZCTAs and counties (a total error rate of less than 0.04%). However, of the 15 un-matched ZIP codes, population is recorded as zero in all of them, and none were found to be matched

in any of the HUD databases. Upon further inspection, several of the ZIP codes are (or were, at one point) in extremely remote areas of Alaska and the southwest where civilization is likely to be zero. For all intents and purposes, the exclusion of these ZIP codes from the weighting procedure is not likely to be problematic.

Appendix D. Variable Time Series Plots

D.1 Operational Variables

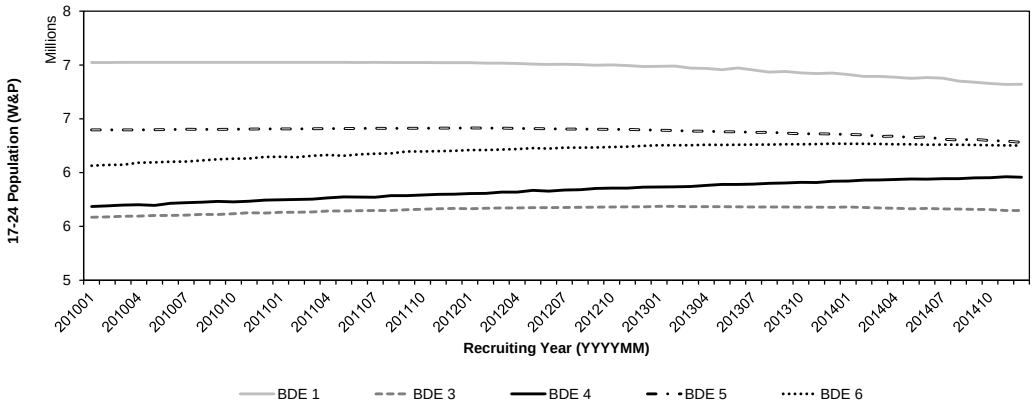


Figure D.1. 17 to 24 Year-Old Population (Source: Woods & Poole, Inc.)

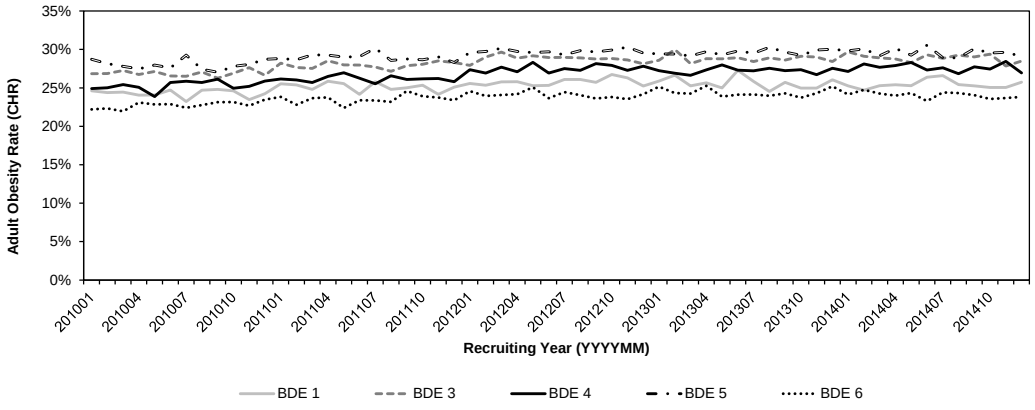


Figure D.2. Adult Obesity Rate (Source: County Health Rankings)

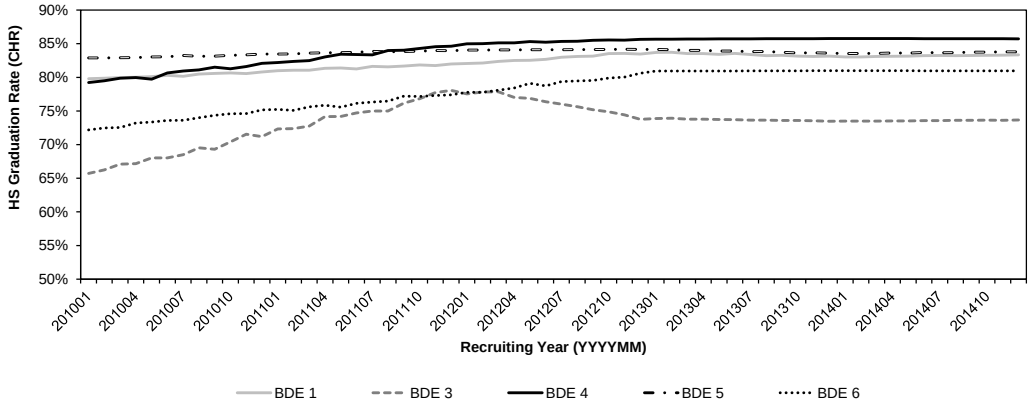


Figure D.3. High School Graduation Rate (Source: County Health Rankings)

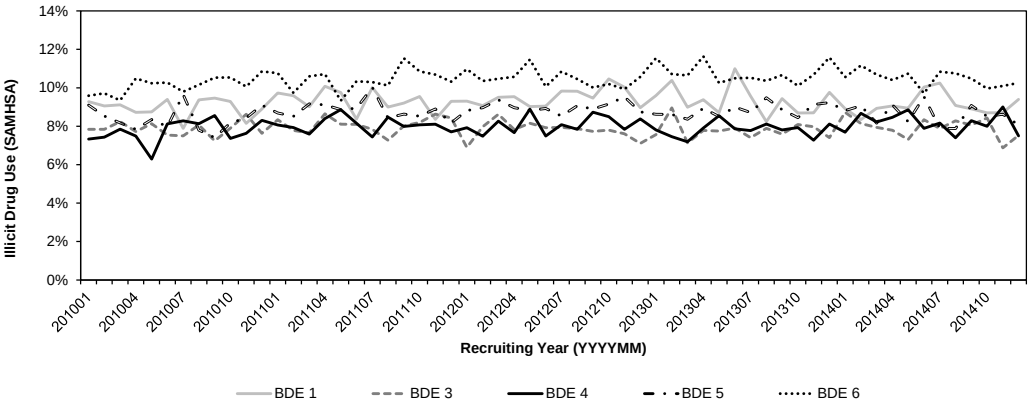


Figure D.4. Illicit Drug Use Rate (Source: County Health Rankings)

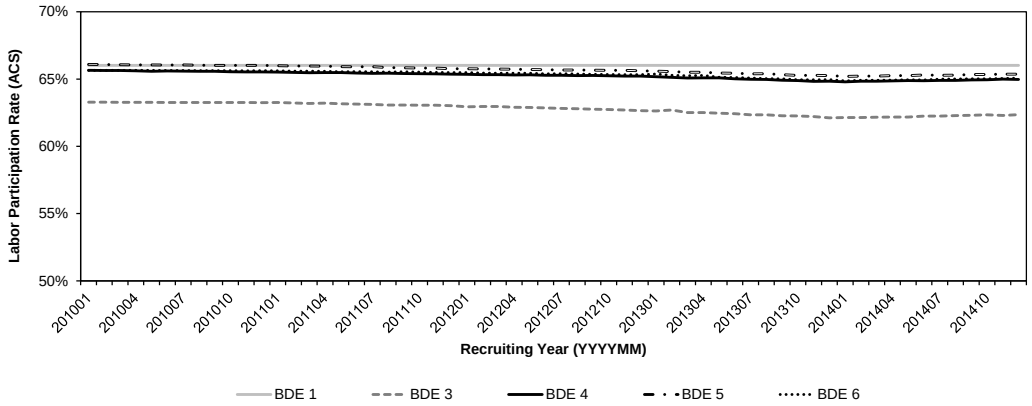


Figure D.5. Labor Participation Rate (Source: 5-Year ACS)

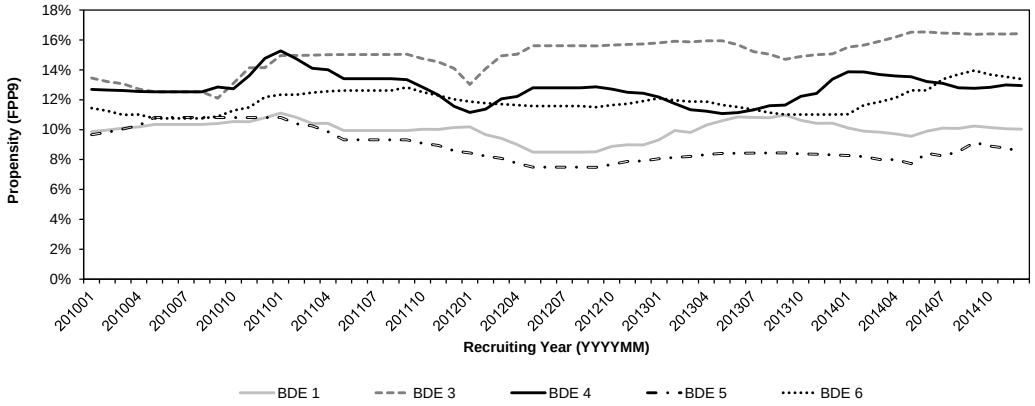


Figure D.6. Propensity (Source: USAREC)

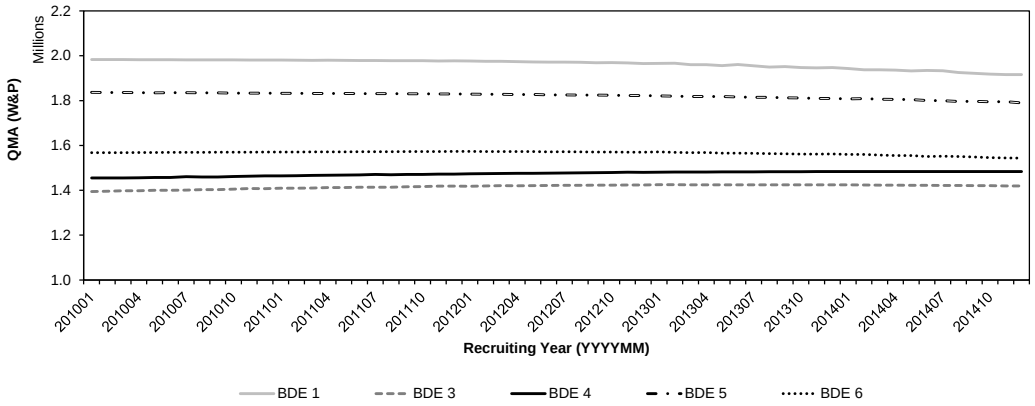


Figure D.7. QMA Population (Source: Woods & Poole, Inc.)

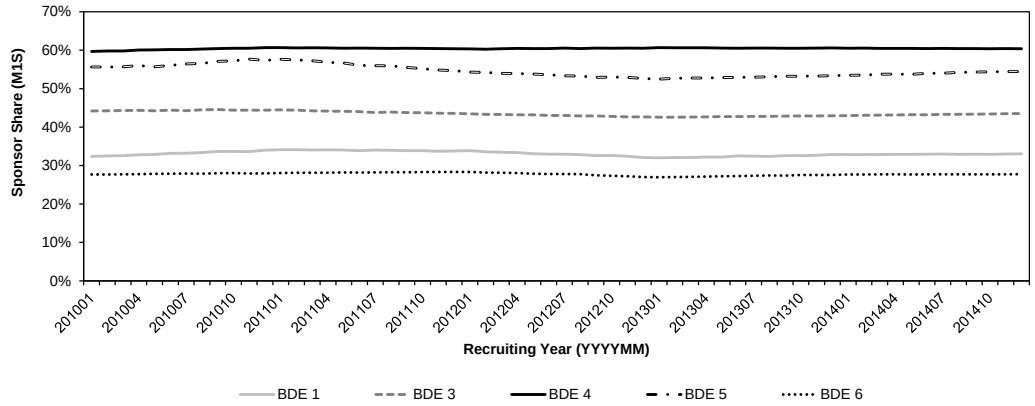


Figure D.8. Sponsor Share (Source: Military One Source)

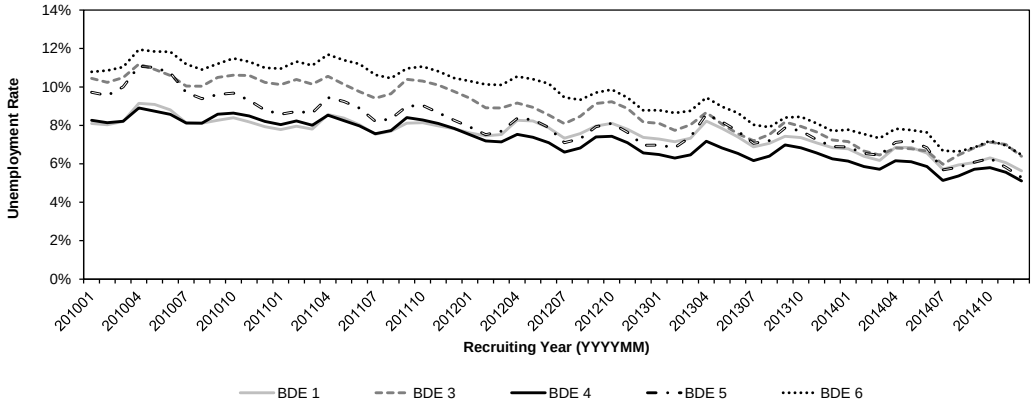


Figure D.9. Unemployment Rate, Not Seasonally Adjusted (Source: LAUS)

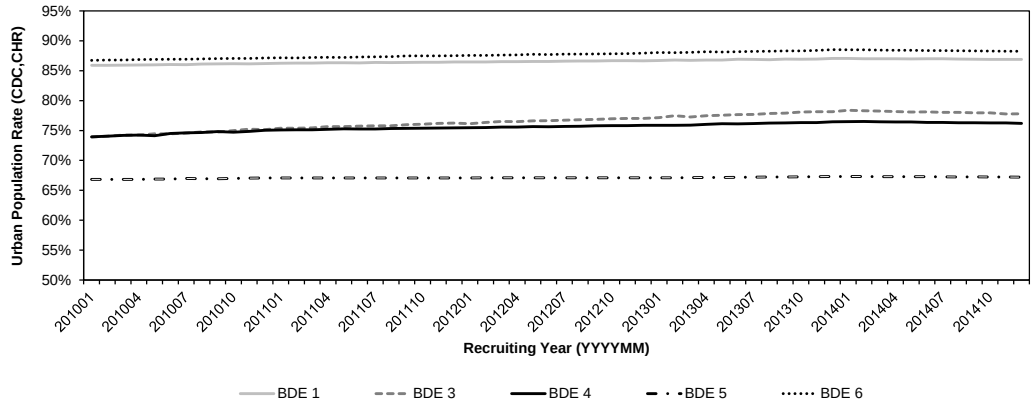


Figure D.10. Proportion of Population Living in Urban Areas (Source: LAUS)

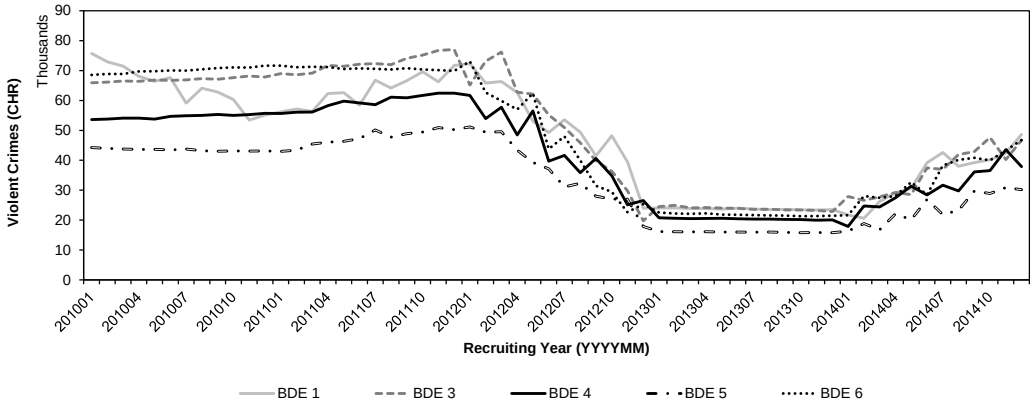


Figure D.11. Violent Crimes (Source: County Health Rankings)

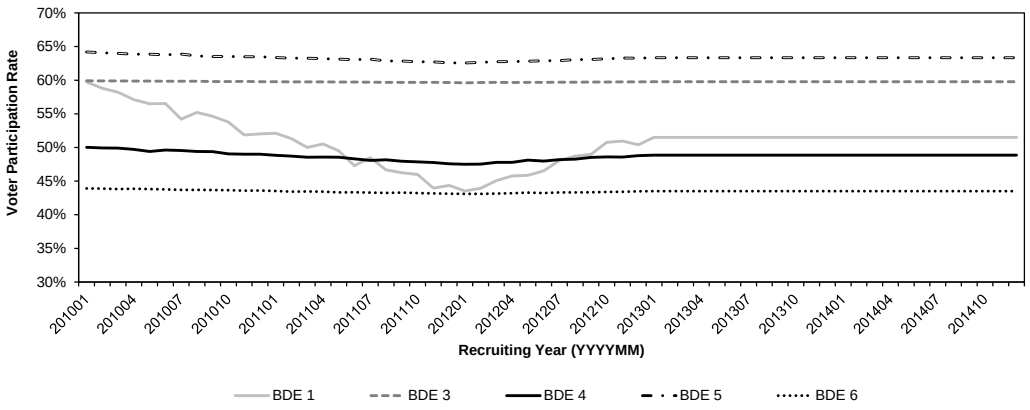


Figure D.12. Voter Participation Rate (Source: *The Guardian*)

D.2 Mission Variables

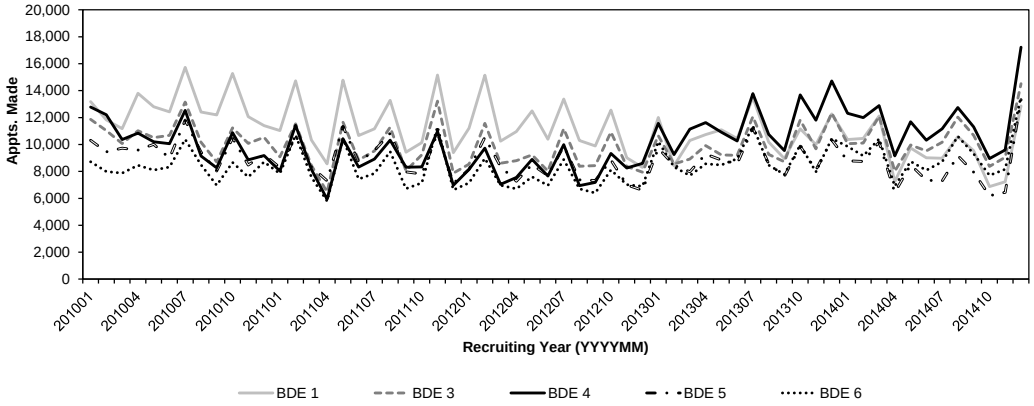


Figure D.13. Appointments Made (Source: USAREC)

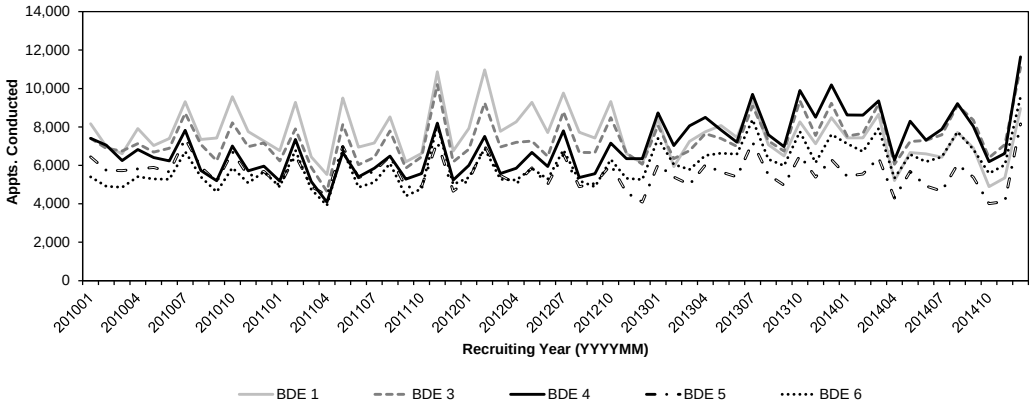


Figure D.14. Appointments Conducted (Source: USAREC)

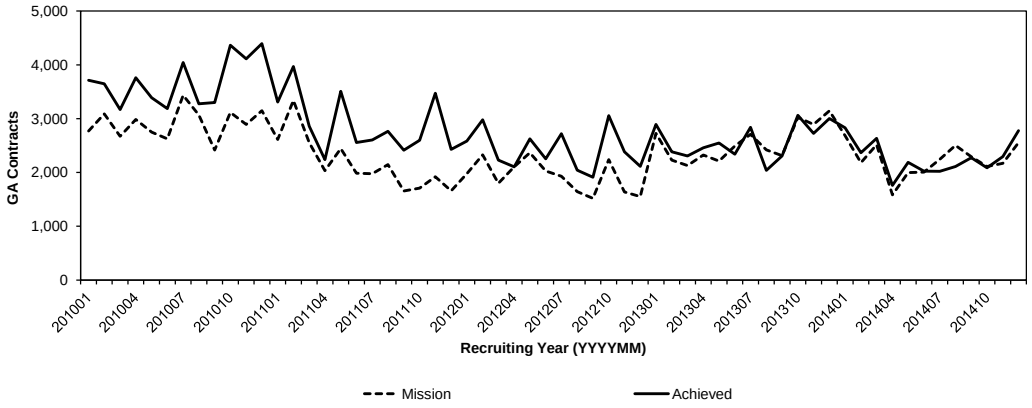


Figure D.15. Graduate Alpha (GA) Contracts (Source: USAREC)

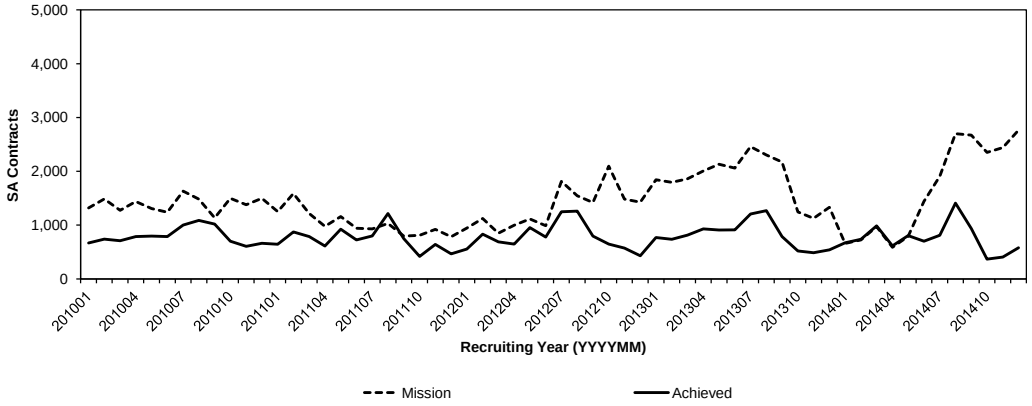


Figure D.16. Senior Alpha (SA) Contracts (Source: USAREC)

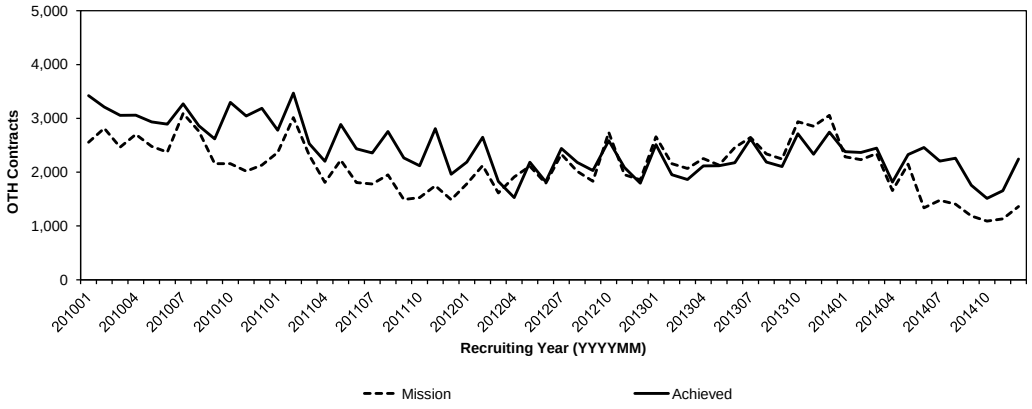


Figure D.17. Other (OTH) Contracts (Source: USAREC)

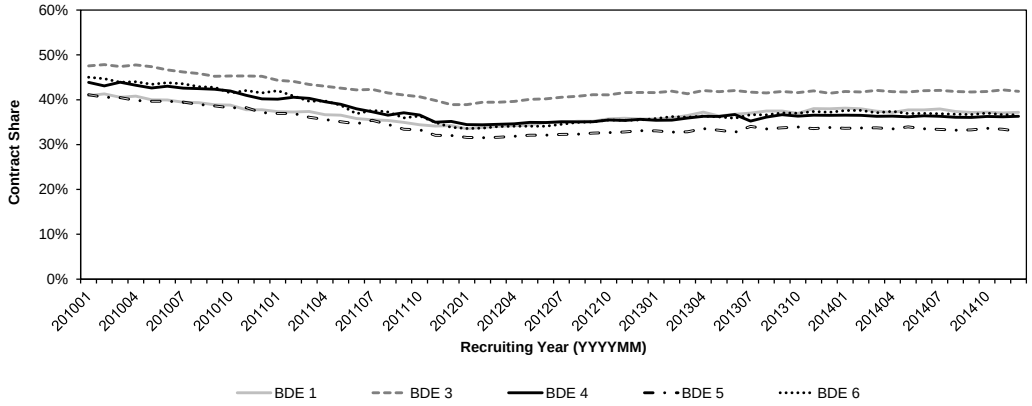


Figure D.18. Contract Share (Source: DMDC)

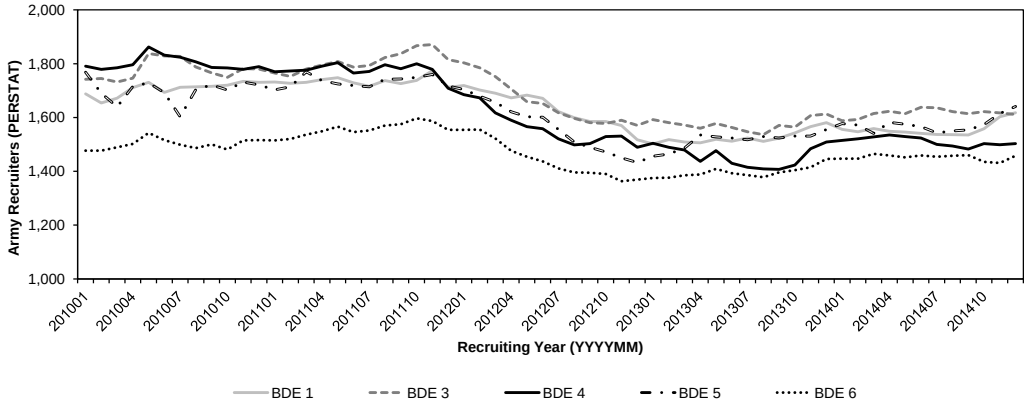


Figure D.19. Army Recruiters (Source: USAREC)

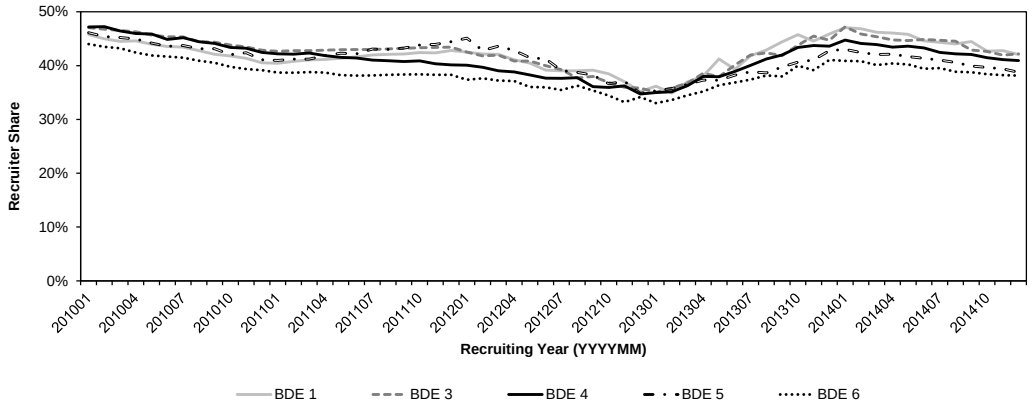


Figure D.20. Recruiter Share (Source: DMDC)

Appendix E. Supplementary Computer Code

E.1 County-to-Battalion Weighting (Microsoft Excel®2010, VBA)

```
Sub WeightCountyData()

Dim rngFIPSCol As Range, rngNumerCol As Range, rngDenomCol As Range, rngErrorCol As Range, rngTemp As Range
Dim strSeriesType As String, strSeriesName As String, strWorkbookName As String, strSeriesDenom As String
Dim i As Long, t As Long, lngN As Long 'counters (i: general, t:time, N: number of time observations)
Dim varTemp As Variant
Dim kSeries As Long

    Application.ScreenUpdating = False

    'These will be captured as user inputs later
    strWorkbookName = "Labor Participation (ACS5).xlsx" 'where the data is located
    strSeriesName = "Labor Participation Rate (ACS)" 'numerator worksheet location
    strSeriesDenom = "Labor Participation Rate (ACS)" 'denominator worksheet location
    strSeriesType = "Rate" 'Rate or Raw
    lngN = 5 'how many columns (time periods) of input data to weight

'*******************************************************************************************
    'Capture series names (assumes 1 series)
    For i = 1 To 38
        Worksheets("xferBN").Activate
        Cells(2, i).Value = strSeriesName
    Next i

    For i = 1 To 5
        Worksheets("xferBDE").Activate
        Cells(2, i).Value = strSeriesName
    Next i

    'Copy the FIPS Column (no header) from the user's data
    '(assumes same FIPS alignment structure for numerator and denominator)
    Workbooks(strWorkbookName).Worksheets(strSeriesDenom).Activate
        Set rngFIPSCol = Range("A3:A3145")
        rngFIPSCol.Copy
    Workbooks("DSS_v5.xlsm").Worksheets("scratch").Activate
        Range("A2").Select
    With Selection
        .PasteSpecial xlPasteValues
    End With

    For t = 1 To lngN 'N = 60 for monthly data, 5 for annual

       'Copy the Numerator Column from the user's data
        Workbooks(strWorkbookName).Worksheets(strSeriesName).Activate
            Range("B3").Select 'top left cell of the first (leftmost) desired numerator column
            Set rngNumerCol = Range(ActiveCell.Offset(0, t - 1), ActiveCell.Offset(3142, t - 1))
            rngNumerCol.Copy
        Workbooks("DSS_v5.xlsm").Worksheets("scratch").Activate
            Range("B2").Select 'B -> NUMERATOR COLUMN
        With Selection
            .PasteSpecial xlPasteValues
        End With

        If strSeriesType = "Rate" Then 'Copy the Denominator Column... always TRUE for unemployment rates
            Workbooks(strWorkbookName).Worksheets(strSeriesDenom).Activate
                Range("B3150").Select 'top left cell of the first (leftmost) desired denominator column
                Set rngDenomCol = Range(ActiveCell.Offset(0, t - 1), ActiveCell.Offset(3142, t - 1))
                rngDenomCol.Copy
            Workbooks("DSS_v5.xlsm").Worksheets("scratch").Activate
                Range("C2").Select 'C -> DENOMINATOR COLUMN
            With Selection
                .PasteSpecial xlPasteValues
            End With
        Else 'don't need a denominator column
        End If

        'Copy and Paste weighted BDE Numbers for one time observation, t of N
         Workbooks("DSS_v5.xlsm").Worksheets("geocorr").Activate
            ActiveSheet.PivotTables("PivotTable1").PivotCache.Refresh
            If strSeriesType = "Rate" Then
                Set rngTemp = Range("Q6:Q10") 'rates
            Else
                Set rngTemp = Range("S6:S10") 'raw numbers
            End If
            varTemp = WorksheetFunction.Transpose(rngTemp)
         Worksheets("xferBDE").Activate
            Range(Cells(2 + t, 1), Cells(2 + t, 5)) = varTemp
```

```
        'Copy and Paste weighted BN Numbers for one time observation, t of N
         Worksheets("geocorr").Activate
            ActiveSheet.PivotTables("PivotTable2").PivotCache.Refresh
            If strSeriesType = "Rate" Then
                Set rngTemp = Range("Q15:Q52") 'rates
            Else
                Set rngTemp = Range("S15:S52") 'raw numbers
            End If
            varTemp = WorksheetFunction.Transpose(rngTemp)
         Worksheets("xferBN").Activate
            Range(Cells(2 + t, 1), Cells(2 + t, 38)) = varTemp

    Next t

    Application.ScreenUpdating = True

End Sub
```

E.2 Stochastic Mean Value Imputation (Microsoft Excel®2010, VBA)

```
Sub ParseInterpolate()

Dim lngNumPeriods As Long, j As Long, lngMonthsPerPd As Long, k As Long, lngStdErr As Long, lngRandErr As Long
Dim dblPeriodPointEst As Double, lngErrorMargin As Long, dblRand As Double, lngRandError As Double, dblStdDev As Double
Dim dblStdError As Double
Dim lngNumUnitCounties As Long, dblPeriodAvg As Double, dblRange As Double, dblRangeStep As Double, dblImputedLast As Double
Dim arrTransferArray As Variant, dblPeriodAvgNext As Double, strIsCumulative As String, strNeedsFinishVal As String
Dim strTypeData As String

    lngNumVars = 1 'adjust as required
    lngNumPeriods = 5 'adjust as required (ANNUAL = 5 for 2010 to 2014)
    lngMonthsPerPd = 12 'adjust as required (*12 for annual, 6 for semi annual, 3 for quarterly data)
    strIsCumulative = "N" 'is the data cumulative? (Y/N)
    strNeedsFinishVal = "N" 'does the data need a final value in order to interpolate the last year's monthly values?
    strTypeData = "Decimal" 'what type of data needs to be interpolated? (Integer, Decimal)
    '********lngNumPeriods*lngMonthsPerPd = 60*************

    Application.ScreenUpdating = False
    For Each sht In ActiveWorkbook.Sheets
        sht.Activate
        ActiveSheet.Unprotect
    Next

    For lngUnit = 1 To 38

        For i = 1 To lngNumVars 'number of cells (variables) to parse per Unit (BDE or BN)
            strUnitID = Worksheets("xferBN").Cells(1, lngNumVars * lngUnit - lngNumVars + i)
            lngColPasteTo = Worksheets(strUnitID).Cells(1, Columns.Count).End(xlToLeft).Column + 1
            Worksheets("xferBN").Range("A2").Offset(0, lngNumVars * lngUnit - lngNumVars + i - 1).Copy
            Worksheets(strUnitID).Activate
            Worksheets(strUnitID).Cells(1, lngColPasteTo).Select
                With Selection
                    .PasteSpecial xlPasteValues
                End With

            Worksheets("xferBN").Activate

            If strNeedsFinishVal = "Y" Then
                Set arrTransferArray = Worksheets("xferBN").Range("A2:A" & lngNumPeriods + 2)
                                            .Offset(0, lngNumVars * lngUnit - lngNumVars + i - 1)
                dblImputedLast = WorksheetFunction.Round(WorksheetFunction.Average(arrTransferArray), 0)
                Cells(lngNumPeriods + 3, lngNumVars * lngUnit - lngNumVars + i) = dblImputedLast
            Else
            End If

            For j = 1 To lngNumPeriods 'for each period (row) of data in the BDE/BN XFER sheet
                If strIsCumulative = "Y" Then 'if the data is a cumulative total, divide by the number of periods
                    dblPeriodAvg = Worksheets("xferBN").Cells(j + 2, lngNumVars * lngUnit - lngNumVars + i).Value / lngMonthsPerPd
                    dblPeriodAvgNext = Worksheets("xferBN").Cells(j + 3, lngNumVars * lngUnit - lngNumVars + i).Value / lngMonthsPerPd
                Else 'if the data is not cumulative, use the given value as the point estimate
                    dblPeriodAvg = Worksheets("xferBN").Cells(j + 2, lngNumVars * lngUnit - lngNumVars + i).Value
                    dblPeriodAvgNext = Worksheets("xferBN").Cells(j + 3, lngNumVars * lngUnit - lngNumVars + i).Value
                End If

                dblRange = dblPeriodAvgNext - dblPeriodAvg
                dblRangeStep = dblRange / lngMonthsPerPd
                dblStdDev = Abs(dblRange / 4) 'assumes +/-2sigma per empirical rule
                Worksheets(strUnitID).Activate

                For k = 1 To lngMonthsPerPd 'for each month in each year, generate a random error about the trend line
                    dblStdError = WorksheetFunction.NormSInv(Rnd) * dblStdDev / (lngMonthsPerPd ^ 0.5)
                    If strTypeData = "Decimal" Then 'if decimal, don't round
                        Worksheets(strUnitID).Cells(1 + j * lngMonthsPerPd - lngMonthsPerPd + k, lngColPasteTo).Value _
                            = dblPeriodAvg + ((k - 1) * dblRangeStep) + dblStdError
                    Else ' it is integer
                        Worksheets(strUnitID).Cells(1 + j * lngMonthsPerPd - lngMonthsPerPd + k, lngColPasteTo).Value _
                            = WorksheetFunction.Round(dblPeriodAvg + ((k - 1) * dblRangeStep) + dblStdError, 0)
                    End If
                Next k
            Next j
        Next i
        Call ProtectSheet
    Next lngUnit
    Application.ScreenUpdating = True
End Sub
```

E.3 Principal Components Analysis (MATLAB®2014)

```
function [EIGVALS_R,EIGVAL_Percent_Var,EIGVAL_CumPercent_Var,L] = mvapca(X,Categories)
%This function completes a Principal Component Analysis on a matrix of any
%size using the correlation matrix.
%   INPUTS:
%       1. X, the data to be analyzed with N observations and p variables
%       2. Categories, an N x 1 vector of up to 11 integer categories
%   OUTPUTS:
%       1. A vector of eigenvalues
%       2. A vector of the percent of variance explained by each eigenvalue
%       3. A vector of cumulative percents from (2)
%       4. A plot of Horn's Test (actual data v. Horn's curve)
%       5. A subplot of all RETAINED component scores against each other
%Created by: Joshua McDonald | AFIT Dept. of Operational Sciences | 4/17/15
%**************************************BEGIN MAIN SCRIPT**********************************
%*****************************************************************************************

[obs,~] = size(X);
[X_S,~,R] = mvastandard(X); %standardize data & output the correlation matrix R
[~,variables] = size(R);

[A_R,EIGVALS_R] = eig(R); %get normalized eigenvectors and values

EIGVALS_R = diag(EIGVALS_R,0)'; %put eigenvalues in a row vector for sorting in decreasing order
for i = 1:(variables-1);
    [~,index] = max(EIGVALS_R(1,i:variables));
    % Swap Values
    moved_val = EIGVALS_R(:,i);
    EIGVALS_R(:,i) = EIGVALS_R(:,i-1+index);
    EIGVALS_R(:,i-1+index) = moved_val;%<-- Sorted Eigenvalues
    % Swap Vectors
    moved_vec = A_R(:,i);
    A_R(:,i) = A_R(:,i-1+index);
    A_R(:,i-1+index) = moved_vec;
end

for i = 1:size(EIGVALS_R,2) %get the percent of variance provided by each eigenvalue
    EIGVAL_Percent_Var(1,i) = EIGVALS_R(1,i)/sum(EIGVALS_R);%<-- Eigenvalue percent variance
end

EIGVAL_CumPercent_Var = cumsum(EIGVAL_Percent_Var) %<-- Cumulative Eigenvalue variance
Y_R = X_S*A_R; %<-- Component Scores (N x p)
L = corr(X_S,Y_R); %<-- Loadings Matrix (p x p)

%****************************HORN'S CURVE***********************************
for i = 1:length(L); %make a vector of component column indices for x-axis
    components(1,i) = i;
end

[curvepoints] = Hornscurve(obs,length(L)); %construct Horn's Curve eigenvalues
function [curvepoints] = Hornscurve(N,p)
%This function takes as INPUTS a number of observations N, and a number of variables p.  It creates
%K=1000 random, NID matricies of size (N x p) and extracts eigenvalues from the Covariance of the
%matrix.  After K rows of p eigenvalues are recorded, the average of the p columns is recorded in a
%1 x p vector, which are the points forming Horn's curve.
%****************Created by: Joshua McDonald | AFIT Dept. of Operational Sciences | 5/5/15

K = 1000;
Eigvals_master = zeros(K,p);

for i = 1:K
    M = randn(N,p);
    C = cov(M);
    [~,Eigvals_C] = eig(C); %get normalized eigenvectors and values
    Eigvals_C = diag(Eigvals_C,0)'; %put eigenvalues in a row vector

    for j = 1:p-1;  % Sort the eignenvalues from largest to smallest
        [~,index] = max(Eigvals_C(1,j:p));
        moved_val = Eigvals_C(:,j);
        Eigvals_C(:,j) = Eigvals_C(:,j-1+index);
        Eigvals_C(:,j-1+index) = moved_val;
    end
        Eigvals_master(i,:) = Eigvals_C(1,:);
end

curvepoints = mean(Eigvals_master);

%*****************************************************************************************
%****************************************END MAIN*****************************************
```

E.4 Durbin-Watson Statistics for Categorical Variables (MATLAB®2014)

```
function [D] = mcdonaldDW(E);
%INPUTS -> E, a T x m matrix of residuals obtained from OLS regression
%   where T (indexed by t) is the number of time observations in each of m
%   columns (indeixed by i)
%OUTPUTS -> D, a m x 1 vector of Durbin-Watson test statistics
%*****************************Created by: J. McDonald, AFIT | 10/17/15


[T,m] = size(E);
e = E;

esumsqr = zeros(1,m);
esqrdiff = zeros(T-1,m);
esumsqrdiffs = zeros(1,m);
D = zeros(1,m);

for i = 1:m
    esumsqr(1,i) = sumsqr(e(:,i));
    for t = 2:T
        esqrdiff(t,i) = (e(t,i)-e(t-1,i))^2;
    end
    esumsqrdiffs(1,i) = sum(esqrdiff(:,i));

    D(1,i) = esumsqrdiffs(1,i)/esumsqr(1,i);
end

D = D';

%*********************************END MAIN**********************************
```

Appendix F. Final Battalion Regression Models

Table F.1. Battalion-echelon Models for Graduate Alpha ($k = GA$) Contracts

i	$\beta_0^{(k,i)}$	$\beta_4^{(k,i)}$	$\beta_{30}^{(k,i)}$	$\beta_{31}^{(k,i)}$	$\beta_{32}^{(k,i)}$	$\beta_{33}^{(k,i)}$	$\phi_{t-1}^{(k,i)}$
BN_1A	6.0013	26.4225	−4.2724	0.0245	-2.0811	1.06×10^{-5}	0.0028
BN_1B	3.1563	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	−0.0010
BN_1D	3.5834	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	−0.0029
BN_1E	2.1978	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0020
BN_1G	6.5452	26.4225	−7.4793	0.0245	-2.0811	1.06×10^{-5}	0.0103
BN_1K	4.7563	26.4225	0.8758	0.0245	-2.0811	-1.85×10^{-5}	−0.0008
BN_1N	2.1978	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0069
BN_1O	4.2881	26.4225	0.8758	0.0245	-2.0811	-3.13×10^{-5}	0.0108
BN_3A	5.6542	26.4225	−3.0809	0.0245	-2.0811	1.06×10^{-5}	−0.0031
BN_3D	3.9949	26.4225	0.8758	0.0245	-2.0811	-3.23×10^{-5}	0.0070
BN_3E	−1.9996	72.9264	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0020
BN_3G	0.8969	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0106
BN_3H	0.8771	26.4225	0.8758	0.0245	2.9382	1.06×10^{-5}	0.0046
BN_3J	2.1978	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0058
BN_3N	0.4258	47.3275	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0043
BN_3T	0.3884	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0199
BN_4C	2.1978	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0069
BN_4D	1.5753	26.4225	0.8758	0.0358	-2.0811	1.06×10^{-5}	0.0021
BN_4E	3.3289	26.4225	0.8758	0.0245	-2.0811	-6.42×10^{-6}	0.0060
BN_4G	2.1978	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0087
BN_4J	2.1978	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0074
BN_4K	−1.0958	60.5243	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0128
BN_4P	2.6972	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0023
BN_5A	1.3459	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0094
BN_5C	4.1218	4.2068	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0022
BN_5D	2.1978	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0045
BN_5H	2.1978	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0060
BN_5I	−0.0926	41.0251	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0069
BN_5J	1.4253	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0151
BN_5K	2.1978	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0104
BN_5N*	2.1978	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0058
BN_6F	4.9470	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	−0.0054
BN_6H	2.1978	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0078
BN_6I	3.7050	26.4225	0.8758	0.0245	-2.0811	-7.31×10^{-6}	−0.0056
BN_6J	1.1932	26.4225	0.8758	0.0437	-2.0811	1.06×10^{-5}	−0.0005
BN_6K	3.7902	26.4225	0.8758	0.0245	-2.0811	-1.02×10^{-5}	0.0079
BN_6L	2.4804	26.4225	0.8758	0.0245	-2.0811	1.06×10^{-5}	0.0089
BN_6N	3.4910	26.4225	−3.7986	0.0245	-2.0811	-1.73×10^{-5}	0.0034

*Baseline

Table F.2. Battalion-echelon Models for Other ($k = OTH$) Contracts

i	$\hat{\beta}_0^{(k,i)}$	$\hat{\beta}_4^{(k,i)}$	$\hat{\beta}_{30}^{(k,i)}$	$\hat{\beta}_{31}^{(k,i)}$	$\hat{\beta}_{32}^{(k,i)}$	$\hat{\beta}_{33}^{(k,i)}$	$\hat{\phi}_{t-1}^{(k,i)}$
BN_1A	3.7519	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0040
BN_1B	4.8331	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	−0.0009
BN_1D	3.3435	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0146
BN_1E	3.0128	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0185
BN_1G	4.8331	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	−0.0039
BN_1K	5.7101	15.2018	-0.9503	0.0200	-0.2294	-2.26×10^{-5}	0.0146
BN_1N	3.7381	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	−0.0040
BN_1O	4.4737	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0193
BN_3A	4.9531	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0056
BN_3D	5.6615	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	−0.0008
BN_3E	−0.5804	63.6241	-0.9503	0.0200	-0.2294	3.68×10^{-5}	0.0069
BN_3G	8.3778	15.2018	-0.9503	0.0200	-0.2294	-4.67×10^{-5}	−0.0015
BN_3H	4.2080	15.2018	-0.9503	0.0200	3.1698	4.82×10^{-6}	0.0080
BN_3J	4.6361	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0099
BN_3N	−0.0017	39.9493	-0.9503	0.0200	-0.2294	4.95×10^{-5}	0.0016
BN_3T	8.0054	15.2018	-4.3841	0.0200	-0.2294	4.82×10^{-6}	0.0029
BN_4C	4.3254	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0066
BN_4D	3.4651	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0146
BN_4E	4.8331	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	−0.0006
BN_4G	3.5428	15.2018	-0.9503	0.0298	-0.2294	4.82×10^{-6}	−0.0004
BN_4J	4.7788	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0079
BN_4K	0.5943	64.6082	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0086
BN_4P	2.7029	50.8070	-0.9503	0.0200	-0.2294	4.82×10^{-6}	−0.0055
BN_5A	3.4791	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0074
BN_5C	4.2807	15.2018	-0.9503	0.0211	-0.2294	4.82×10^{-6}	−0.0078
BN_5D	4.0023	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0060
BN_5H	6.4252	15.2018	-0.9503	0.0114	-0.2294	4.82×10^{-6}	−0.0157
BN_5I	4.4255	15.2018	-0.9503	0.0200	-6.0886	4.82×10^{-6}	0.0079
BN_5J	3.2492	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0115
BN_5K	3.3761	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0106
BN_5N*	4.8331	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	−0.0015
BN_6F	4.3130	15.2018	-0.9503	0.0200	-0.2294	-1.71×10^{-5}	0.0225
BN_6H	4.2229	15.2018	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0163
BN_6I	4.8073	15.2018	-0.9503	0.0129	-0.2294	4.82×10^{-6}	0.0044
BN_6J	0.9208	40.7777	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0190
BN_6K	−0.6777	50.6793	-0.9503	0.0200	-0.2294	4.82×10^{-6}	0.0185
BN_6L	4.0834	15.2018	-0.9503	0.0200	-4.1426	4.82×10^{-6}	0.0160
BN_6N	−0.0084	61.7380	-0.9503	0.0200	-0.2294	-7.84×10^{-6}	0.0059

*Baseline

Table F.3. Battalion-echelon Models for Other ($k = SA$) Contracts

i	QTR 1* $\hat{\beta}_0^{(k,i)}$	QTR 2 $\hat{\beta}_0^{(k,i)}$	QTR 3 $\hat{\beta}_0^{(k,i)}$	QTR 4 $\hat{\beta}_0^{(k,i)}$	$\hat{\beta}_4^{(k,i)}$	$\hat{\beta}_{30}^{(k,i)}$	$\hat{\beta}_{31}^{(k,i)}$	$\hat{\beta}_{33}^{(k,i)}$	$\hat{\phi}_{t-1}^{(k,i)}$
BN_1A	3.2325	3.5766	4.0698	2.8806	-12.7345	0.5186	0.0107	1.47×10^{-6}	0.0051
BN_1B	-2.1747	-1.8306	-1.3374	-2.5265	-12.7345	7.3834	0.0107	1.47×10^{-6}	0.0155
BN_1D	2.9588	3.3029	3.7961	2.6070	-12.7345	0.5186	0.0107	1.47×10^{-6}	0.0335
BN_1E	3.2325	3.5766	4.0698	2.8806	-12.7345	0.5186	0.0107	1.47×10^{-6}	0.0004
BN_1G	1.8122	2.1564	2.2980	1.4604	-12.7345	0.5186	0.0107	1.47×10^{-6}	0.0314
BN_1K	3.2325	3.5766	4.0698	2.8806	-12.7345	0.5186	0.0107	8.82×10^{-6}	0.0314
BN_1N	3.6116	3.9557	3.5476	3.8873	-12.7345	0.5186	0.0107	1.47×10^{-6}	-0.0053
BN_1O	-2.2302	-1.2794	-1.3929	-2.5820	-12.7345	6.8295	0.0107	1.47×10^{-6}	0.0185
BN_3A	9.5942	9.9383	10.4315	9.2424	-69.8499	0.5186	0.0107	1.47×10^{-6}	-0.0159
BN_3D	-1.6317	-1.2875	-0.7943	-1.9835	-12.7345	0.5186	0.0107	1.30×10^{-4}	-0.0065
BN_3E	0.6327	0.9769	1.4701	0.2809	-12.7345	5.3432	0.0107	1.47×10^{-6}	0.0126
BN_3G	3.7872	4.1313	4.6245	3.4354	-12.7345	0.5186	0.0023	1.47×10^{-6}	0.0179
BN_3H	3.2325	3.5766	4.0698	2.8806	-12.7345	0.5186	0.0107	1.47×10^{-6}	0.0021
BN_3J	6.9493	7.2935	7.7866	6.5975	-44.7603	0.5186	0.0107	1.47×10^{-6}	0.0015
BN_3N	1.9630	2.3071	2.8003	0.9800	-12.7345	4.6567	0.0107	1.47×10^{-6}	0.0032
BN_3T	3.4437	3.7878	4.2810	3.0919	-12.7345	0.5186	0.0107	1.47×10^{-6}	-0.0221
BN_4C	8.1182	8.4624	8.9555	7.7664	-12.7345	-4.5569	0.0107	1.47×10^{-6}	-0.0141
BN_4D	6.8775	7.2216	7.7148	6.5257	-49.2110	0.5186	0.0107	1.47×10^{-6}	-0.0095
BN_4E	3.4288	3.7730	4.2662	2.5276	-12.7345	0.5186	0.0107	1.47×10^{-6}	0.0082
BN_4G	3.7728	4.1170	4.6102	3.4210	-12.7345	0.5186	0.0107	1.47×10^{-6}	-0.0131
BN_4J	3.6719	3.4995	4.5093	3.3201	-12.7345	0.5186	0.0107	1.47×10^{-6}	0.0076
BN_4K	3.2325	3.5766	4.0698	2.8806	-12.7345	0.5186	0.0107	1.47×10^{-6}	0.0037
BN_4P	6.3981	6.2528	7.2354	6.0463	-25.3920	0.5186	0.0107	1.47×10^{-6}	-0.0248
BN_5A	3.2325	3.5766	4.0698	2.8806	-12.7345	0.5186	0.0107	1.47×10^{-6}	-0.0126
BN_5C	5.6229	5.9670	6.4602	5.2711	-12.7345	-2.6684	0.0107	1.47×10^{-6}	0.0042
BN_5D	3.2325	3.5766	4.0698	2.8806	-12.7345	0.5186	0.0107	1.47×10^{-6}	-0.0052
BN_5H	2.9113	3.2555	3.7487	2.5595	-12.7345	0.5186	0.0192	1.47×10^{-6}	0.0128
BN_5I	3.2325	3.5766	4.0698	2.8806	-12.7345	0.5186	0.0107	1.47×10^{-6}	0.0008
BN_5J	6.2880	6.6322	7.1253	5.9362	-12.7345	-3.0345	0.0107	1.47×10^{-6}	0.0180
BN_5K	3.0773	3.0012	3.9147	2.7255	-12.7345	0.5186	0.0107	1.35×10^{-5}	0.0042
BN_5N*	3.2325	3.5766	4.0698	2.8806	-12.7345	0.5186	0.0107	1.47×10^{-6}	0.0030
BN_6F	3.2325	3.5766	4.0698	2.8806	-12.7345	0.5186	0.0107	1.47×10^{-6}	-0.0046
BN_6H	3.6321	3.9763	4.4695	3.2803	-12.7345	0.5186	0.0107	-1.01×10^{-5}	0.0137
BN_6I	3.5007	3.8449	4.3380	3.1489	-12.7345	0.5186	0.0160	1.47×10^{-6}	0.0061
BN_6J	0.2422	0.5864	1.0795	-0.1096	26.5042	0.5186	0.0107	1.47×10^{-6}	0.0145
BN_6K	3.6321	3.9763	4.4695	3.2803	-12.7345	2.6420	0.0107	-1.01×10^{-5}	0.0145
BN_6L	3.5736	3.9177	4.4109	3.2218	-12.7345	0.5186	0.0107	1.47×10^{-6}	0.0209
BN_6N	2.6417	2.9858	3.4790	2.2899	15.0886	0.5186	0.0107	-2.53×10^{-5}	-0.0131

*Baseline

Analysis and Modeling of U.S. Army Recruiting Markets

Sponsor: Mr. Joe Baird
HQ USAREC/G-2
Ft. Knox, KY

MAJ Joshua Mc Donald
Advisor: Dr. Edward D. White
Department of Mathematics and Statistics (ENC)
Air Force Institute of Technology

PURPOSE

Provide US Army Recruiting Command (USAREC) leadership with focused, relevant, and quantitative insight into its missioning process. *Missioning* is the process whereby missions (recruiting equivalents of a sales goals) are assigned to subordinate recruiting units in order to maximize the number of enlistments produced.

RESEARCH QUESTION

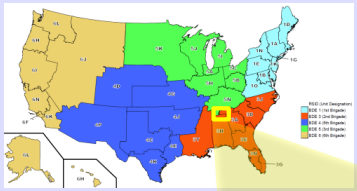
To what extent can we accurately express the relationship between enlistment supply[1] and demand[2] factors, and enlistment contract production?

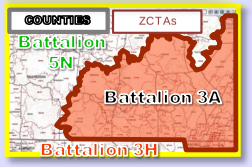
BACKGROUND

MOTIVATION

Previous efforts to model behaviors of recruiting markets leave two primary knowledge gaps:

(1) Market boundaries which are not directly applicable to the 38 current USAREC unit boundaries (shown):





- USAREC unit boundaries are defined by ZIP Code Tabulation Areas (ZCTAs) whereas reliable data exists only and the county-and-higher level
- Previous studies have not adequately accounted for this incongruity

(2) No use of validation datasets to thoroughly test the stability of enlistment production models, or provide estimates of response uncertainty in the future

[1]Recruiting supply factors are outside of USAREC's control
[2]Recruiting demand factors are within USAREC's control
[3]High aptitude high school seniors (SA), high aptitude high school graduates (GA), and all others (OTH)

METHODOLOGY

1 Data Collection

- **Operational Variables (PMESII-PT)** and corresponding sub-variables describe aspects mostly of recruiting supply factors
- **Mission Variables (METT-TC)** describe salient aspects of recruiting demand factors
- **26 total metrics represent both types of variables** from county-level open sources and USAREC-internal databases
  - **County-level data weighted to ZCTAs** then aggregated to recruiting battalion levels with

$$z'_i = \sum_{m \in Z_i} \sum_{n \in C_i} v_{m(n)} z_n$$

Set of ZCTAs defining battalion i | Set of counties which intersect Z_i

where

$v_{m(n)} \equiv$ the proportional population of county n residing in ZCTA m, as of 2010 Census

$z_n \equiv$ the available datapoint for county n

  - **Stochastic Mean Value Imputation** applied to create monthly data points from gaps in available annual data

2 Variance Reduction

- **Principal Components Analysis** used to reduce multicollinearity between the inter-related aspects of the collected data
  - **Horn's criteria** helps determine how many principal components of variance are useful to keep

3 Estimation: Stepwise Ordinary Least Squares

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

- Stepwise methods estimate the OLS model to optimize fit (i.e., values of $R^2_{Adjusted}$); parsimony is obtained by iteravely entering statistically significant terms (α_{IN} = 0.05) and removing non-significant terms (α_{OUT} = 0.1)
  - **Categorical variables** model recruiting units and quarters
  - **Studentized residuals**, r_i, account for x-space distance
  - Lag-1 **autocorrelations** evaluated with Durbin-Watson test
  - **Multicollinearity** assessed for Variance Inflation Factors (VIFs) > 10
  - Approx. **Box-Cox transformations** correct variance heteroscedasticity

3 Validation

- **25% of data set aside** to test OLS models' predictive accuracy
- Prediction Intervals **account for forecast inputs**

$$100(1-\alpha)\%\,\mathrm{PI} = \hat{y}_{t+1} \pm z_{\alpha/2}\left[\hat{\sigma}_u^2 + \sum_{x_j \in \mathbf{x}} \hat{\beta}_j^2 \hat{\sigma}_j^2\right]^{1/2} \qquad t = T, T+1, \ldots, T+\tau-1$$

- Validation metrics:

$$MAPE = 100\% \cdot \frac{1}{N}\sum_{t=T+1}^{T+\tau}\left|\frac{y_t - \hat{y}_t}{y_t}\right|; \quad MAD = \frac{1}{N}\sum_{t=T+1}^{T+\tau}\left|y_t - \hat{y}_t\right|; \quad RMSE = \left[\frac{1}{N}\sum_{t=T+1}^{T+\tau}\left(y_t - \hat{y}_t\right)^2\right]^{1/2}$$

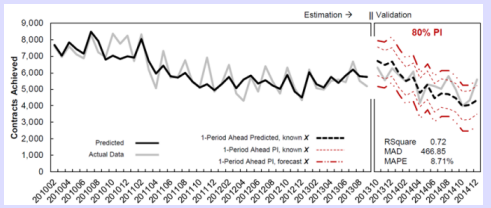
RESULTS & ANALYSIS

Key Takeaway#1

- A set of 5 continuous variables + 40 categorical variables for battalions and quarters **explains 70%, 74%, and 81% of the estimation data for SA, GA, and OTH enlistment contracts[3]**

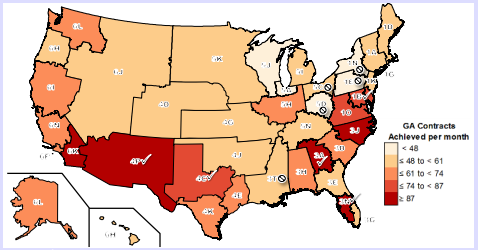
Key Takeaway#2

- Validation shows models are stable even with forecast inputs



Key Takeaway#3

- Markets can be individually characterized according to predictive accuracy; **Example: GA**



✓ **Top-5** per month production and better-than average degradation in % accuracy
⊘ **Bottom-5** per month production and better-than average degradation in % accuracy

CONCLUSIONS

- We achieve **530%, 170%, and 119% relative increases over previous efforts** to model SA, GA, and OTH enlistment contract production
- Market-specific models will likely **allow USAREC to complete the optimization portion of its missioning process more effectively**
- **Markets at the extremes of production and having poor accuracy in validation warrant further investigation**

Department of Mathematics and Statistics

Bibliography

1. Headquarters: Department of the Army, Washington, D.C., *Army Doctrine Reference Publication* (ADRP) *5-0: The Operations Process*, 2012.

2. U.S. Census Bureau, "ZIP Code Tabulation Areas (ZCTAs) Delineation Animation." `http://www.census.gov/geo/reference/zcta/zcta_delin_anim.html`, 2015 [Accessed: 29 July 2015].

3. S. Waddell, *History of the Military Art Since 1914*. West Point, New York: Pearson Custom Publishing, 2005.

4. M. Flesichmann and M. Nelson, "Refining Recruiting Mission Allocation Using a Recruiting Market Index." (Presentation to the Army Operations Research Symposium), 2014.

5. Headquarters: United States Army Training and Doctrine Command, Fort Eustis, Virginia, *TRADOC Pam 525-3-7: The U.S. Army Human Dimension Concept*, 2014.

6. The Economist, "Who will fight the next war?." `http://www.economist.com/node/21676778`, 2015 [Accessed: 24 October 2015].

7. Headquarters: United States Army Recruiting Command, Fort Knox, Kentucky, *USAREC Manual 3-0: Recruiting Operations*, 2009.

8. M. Flesichmann, "2-Step Enlisted Recruiting Mission Model." (PowerPoint Presentation to Dr. White and CPT McDonald), 2015.

9. M. Ben-Akiva and S. Lerman, *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, Massachusetts: Massachusetts Institute of Technology Press, 1985.

10. J. Dertouzos, "Recruiter Incentives and Enlistment Supply," tech. rep., RAND Corporation, Santa Monica, California, 1985.

11. M. Murray and L. McDonald, "Recent Recruiting Trends and Their Implications for Models of Enlistment Supply," tech. rep., RAND Corporation, Santa Monica, California, 1999.

12. J. Warner, C. Simon, and D. Payne, "Enlistment Supply in the 1990's: A Study of the Navy College Fund and Other Enlistment Incentive Programs," tech. rep., Defense Manpower Data Center, JAMRS Division, Arlington, Virginia, 2001.

13. J. Dertouzos and S. Garber, "Human Resource Management and Army Recruiting: Analyses of Policy Options," tech. rep., RAND Corporation, Santa Monica, California, 2006.

14. J. Dertouzos and S. Garber, "Performance Evaluation and Army Recruiting," tech. rep., RAND Corporation, Santa Monica, California, 2008.

15. J. Gibson, R. Hermida, J. Luchman, B. Griepentrog, and S. Marsh, "ZIP Code Valuation Study Technical Report," tech. rep., Joint Advertising, Market Research & Studies (JAMRS), Defense Human Resources Activity, Arlington, Virginia, 2011.

16. J. Gibson, J. Luchman, B. Griepentrog, S. Marsh, A. Zucker, and M. Boehmer, "ZIP Code Valuation Study Technical Report: Predicting Army Accessions," tech. rep., Joint Advertising, Market Research & Studies (JAMRS), Defense Human Resources Activity, Arlington, Virginia, 2009.

17. Q. Vuong, "Likelihood Ratio Tests for Model Selection and Non-Tested Hypotheses," *Econometrica*, vol. 57, no. 2, pp. 307–333, 1989.

18. B. Asch, P. Heaton, and B. Savych, "Recruiting Minorities: What Explains Recent Trends in the Army and Navy?," tech. rep., RAND Corporation, Santa Monica, California, 2009.

19. R. Kilburn and J. Klerman, "Enlistment Decisions in the 1990s: Evidence From Individual-Level Data," tech. rep., RAND Corporation, Santa Monica, California, 1999.

20. J. Hosek and C. Peterson, "Reenlistment Bonuses and Retention Behavior," tech. rep., RAND Corporation, Santa Monica, California, 1985.

21. M. Kleykamp, "College, Jobs, or the Military? Enlistment During a Time of War," *Social Science Quarterly*, vol. 87, no. 2, pp. 272–290, 2006.

22. B. Rostker, J. Klerman, and M. Zander-Cotugno, "Recruiting Older Youths: Insights from a New Survey of Army Recruits," tech. rep., RAND Corporation, Santa Monica, California, 2014.

23. T. Henry, K. Dice, and M. Davis, "A Decision Support Tool for Determining Army Enlistment Initiatives," Tech. Rep. 20020319–195, Operations Research Center of Excellence (United States Military Academy, Deptartment of Systems Engineering), West Point, New York, 2001.

24. B. Bicksler and L. Nolan, "Recruiting an All-Volunteer Force: The Need for Sustained Investment in Recruiting Resources—An Update." (Sponsored by the Directorate of Accession Policy, Under Secretary of Defense for Personnel and Readiness), 2009.

25. R. Howard and A. Abbas, *Foundations of Decision Analysis*. Upper Saddle River, New Jersey: Prentice Hall, 2015.

26. U.S. Census Bureau, "Zip Code Tabulation Areas (ZCTAs)." https://www.census.gov/geo/reference/zctas.html, 2015 [Accessed: 23 September 2015].

27. U.S. Census Bureau, "2010 ZCTA to County Relationship File." http://www2.census.gov/geo/docs/maps-data/data/rel/zcta_county_rel_10.txt, 2015 [Accessed: 23 September 2015].

28. U.S. Census Bureau, "Explanation of the 2010 ZCTA to County Relationship File." http://www2.census.gov/geo/docs/maps-data/data/rel/explanation_zcta_county_rel_10.pdf, 2015 [Accessed: 23 September 2015].

29. D. Montgomery, L. Johnson, and J. Gardiner, *Introduction to Time Series Analysis and Forecasting, 2nd Edition*. Hoboken, New Jersey: John Wiley and Sons, Inc., 2015.

30. D. Wackerly, W. Mendenhall, and R. Schaeffer, *Mathematical Statistics with Applications, 7th Edition*. Belmont, California: Brooks-Cole Cengage, 2008.

31. J. Banks, J. Carson, B. Nelson, and D. Nicol, *Discrete Event Simulation, 5th Edition*. Upper Saddle River, New Jersey: Prentice Hall, 2010.

32. The Guardian, "2008 presidential election results by state and county." http://www.theguardian.com/news/datablog/2009/mar/02/us-elections-2008, 2009 [Accessed: 1 September 2015].

33. The Guardian, "Full US 2012 election county-level results to download." http://www.theguardian.com/news/datablog/2012/nov/07/us-2012-election-county-results-download, 2012 [Accessed: 1 September 2015].

34. U.S. Census Bureau, "American Community Survey 5 Year Data (2010 - 2014)." http://www.census.gov/data/developers/data-sets/acs-survey-5-year-data.html, 2015 [Accessed: 1 September 2015].

35. ICF International, "2010–2013 Demographics Profiles of the Military Community." http://www.militaryonesource.mil/footer?content_id=279104, 2013 [Accessed: 28 September 2015].

36. U.S. Bureau of Labor Statistics, "Local Area Unemployment Statistics (One-Screen Database Search)." http://www.bls.gov/lau/#data, 2015 [Accessed: 6 September 2015].

37. University of Wisconsin Population Health Institute, "County Health Rankings and Roadmaps." http://www.countyhealthrankings.org/rankings/data, 2015 [Accessed: 29 September 2015].

38. National Center for Health Statistics, "NCHS Urban-Rural Classification Scheme for Counties." http://www.cdc.gov/nchs/data_access/urban_rural.htm, 2014 [Accessed: 3 September 2015].

39. C. K. Ricardo Carvalho and S. Marsh, "Department of Defense Youth Poll, Wave 20–December 2010: Overview Report," tech. rep., Joint Advertising, Market Research & Studies (JAMRS), Defense Human Resources Activity, Arlington, Virginia, 2011.

40. A. Pankratz, *Forecasting with Dynamic Regression Models*. New York: John Wiley and Sons, Inc., 1991.

41. W. Dillon and M. Goldstein, *Multivariate Analysis: Methods and Applications*. New York: John Wiley and Sons, Inc., 1984.

42. K. Bauer, "Course Notes, OPER 685: Applied Multivariate Analysis I." (Spring quarter), 2015.

43. D. Montgomery, E. Peck, and G. Vining, *Introduction to Linear Regression Analysis, 5th Edition*. Hoboken, New Jersey: John Wiley and Sons, Inc., 2012.

44. R. Myers, D. Montgomery, and C. Anderson-Cook, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments, 3rd Edition*. Hoboken, New Jersey: John Wiley and Sons, Inc., 2000.

45. B. Bowerman, R. O'Connell, and A. Koehler, *Forecasting, Time Series, and Regression: An Applied Approach*. Belmont: Thomson Brooks/Cole, 2005.

46. S. M. K. Y. Ronald Walpole, Raymond Myers, *Probability & Statistics for Engineers & Scientists, 8th Edition", publisher =*.

47. U.S. Department of Housing and Urban Development, "HUD USPS ZIP Code Crosswalk Files." `http://www.huduser.gov/portal/datasets/usps_crosswalk.html`, 2015 [Accessed: 29 July 2015].

48. U.S. Census Bureau, "Frequently Asked Questions." `https://ask.census.gov/faq.php?id=5000&faqId=10492`, 2015 [Accessed: 29 July 2015].

49. The American Academy of Family Physicians, "ZIP Code to ZCTA Crosswalk." `http://www.udsmapper.org/docs/zip_to_zcta_2015.xlsx`, 2015 [Accessed: 29 July 2015].

50. The American Academy of Family Physicians, "About the UDS Mapper." `http://www.udsmapper.org/about.cfm`, 2015 [Accessed: 29 July 2015].

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
24–03–2016	Master's Thesis	Oct 2014 — Mar 2016

4. TITLE AND SUBTITLE

Analysis and Modeling of U.S. Army Recruiting Markets

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

M^{c}Donald, Joshua L., Major, USA

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/ENC)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENC-MS-16-M-117

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Mr. Joseph Baird
Chief, Market Research Division, USAREC G-2
1307 Third Ave.
Fort Knox, KY 40121
COMM (502) 626-1393
Email: joseph.a.baird4.civ@mail.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The United States Army Recruiting Command (USAREC) is charged with finding, engaging, and ultimately enlisting young Americans for service as Soldiers in the U.S. Army. USAREC must decide how to allocate monthly enlistment goals, by aptitude and education level, across its 38 subordinate recruiting battalions in order to maximize the number of enlistment contracts produced each year. In our research, we model the production of enlistment contracts as a function of recruiting supply and demand factors which vary over the recruiting battalion areas of responsibility. Using county-level data for the period of recruiting year (RY)2010 through RY2013 mapped to recruiting battalion areas, we find that a set of five variables along with categorical indicators for battalions and quarters of the fiscal year accounts for 70%, 74%, and 81% of the variation in contract production for high-aptitude high school seniors, high-aptitude high school graduates and all others, respectively. We find indications that high-aptitude seniors and graduates should be modeled as separate entities, contrary to current procedure. Finally, our models perform consistently well against a validation dataset from RY2014, and we ultimately achieve 530%, 119%, and 170% relative increases in respective correlation coefficients over previous comparable literature.

15. SUBJECT TERMS

Army, recruiting, market analysis, regression, time series, cross-sectional data, forecasting, principal components

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Edward D. White (ENC)
U	U	U	UU	155	19b. TELEPHONE NUMBER *(include area code)* (937)255-3636 x4540 Edward.white@afit.edu